CHILD SPEECH UNDERSTANDING AND GENERATION VIA NEURAL ASR AND TTS MODELS

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Rishabh Jain (19231092)

This dissertation is submitted in fulfillment of the requirement for the degree of Doctor of Philosophy.

School of Electrical and Electronics Engineering
University of Galway

Supervisor:	Dr. Peter Corcoran
Date of Submission:	2nd April 2024

Declaration of Academic Integrity

I hereby confirm that the present thesis is solely my own work and that if any text passages or diagrams from books, papers, the Web or other sources have been copied or in any other way used, all references—including those found in electronic media—have been acknowledged and fully cited.

2nd-April-2024 Rishabh Jain

Acknowledgment

“Life is unpredictable. Not everything's in our control. But as long as you're with the right people, you can handle anything”. This profound quote, which resonated with me while watching Brooklyn Nine-Nine one late night, has remained etched in my memory ever since. Therefore, I would like to express my heartfelt gratitude to those 'right' people in my life who have supported me through thick and thin.

First and foremost, I extend my deepest gratitude to my supervisor and mentor, Professor Peter Corcoran. It was Professor Corcoran who not only recognized my potential but also encouraged me to embark on the journey of pursuing a PhD at the University of Galway. His unwavering support and guidance have been instrumental in shaping me both personally and academically. I am forever thankful for the myriad opportunities he has provided me. His belief in my abilities has been a constant source of motivation, and his mentorship has instilled in me the confidence to navigate through the challenges of the PhD journey. I treasure the insightful discussions and brainstorming sessions we have shared. I also cherish the memory of when Professor Corcoran went above and beyond his duties by personally delivering my computer to my home during the challenging times of the COVID-19 pandemic. It is gestures like these that exemplify his unwavering dedication to his students' well-being and success. In many ways, Professor Corcoran has become more than just a supervisor to me; he is akin to an "Irish Dad," as affectionately coined by my friends. Thank you, Professor Corcoran, for not only being an exceptional supervisor but also a remarkable individual from whom I continue to learn and grow.

Furthermore, I would like to extend my heartfelt thanks to the members of my research group and fellow PhD students. Muhammad Ali Farooq, Andrei Barcovschi, Mariam Yiwere and Waseem Shariff have been exceptionally supportive throughout my PhD journey, always ready to lend a helping hand and provide valuable insights. Special mention goes to Dan Bigioi, whose friendship and support have been invaluable since the beginning of my PhD. Thank you for working and co-authoring papers with me and for the memorable Romanian wedding experience. Your support made my transition into my PhD work seamless.

I must also express my gratitude to Claudia Costache, whose exceptional management skills and unwavering support in administrative matters have been indispensable. Joe Desbonnet deserves recognition for his exemplary problem-solving abilities and willingness to assist, even during late hours, with any technical queries. Special thanks to Michael Schukat, my mentor during my Masters, who introduced me to the world of research. Your guidance and support laid the foundation for my academic journey.

I would also like to express my sincere appreciation to Gabriel Costache from Xperi for his invaluable expertise and guidance during the course of my research. Despite his demanding role as a senior manager at Xperi and team lead for the DAVID project, Gabriel consistently made time to address my queries and offer his insights. His kind and humble nature left a lasting impression on me and his qualities of expertise, kindness, and humility are admirable traits that I aspire to emulate in my own journey.

I am also grateful to the researchers and engineers I had the privilege of collaborating with at Xperi on the DAVID project. Francisco Salgado, George Sterpu, and Sathish Mangapuram deserve special mention for their technical expertise and innovative solutions, which greatly contributed to the advancement of my research. I extend my gratitude to Joe Lemley for his valuable advice and insights, despite our limited collaboration during my PhD.

Additionally, I extend my thanks to the Xperi-USA team, particularly Zoran Ferzo, for his expertise and support in my research endeavors, as well as to the rest of the team members for their collaboration and assistance.

I am also grateful for the friendships forged and the kindness shown by my colleagues at Xperi. Victor Vlad, Ayush Rai, Luke Connolly, Bentolhoda Binaei, Samira Pourkhajouei, Mohamed Moustafa, Francisco Raposo, Padraic Toomey, Shubhajit Basak, and others have made my time at Xperi truly enjoyable and enriching.

Furthermore, I will always be grateful for my colleagues and friends at Fotonation/Xperi, University of Galway, and the C3 Imaging lab.

I would also like to thank Horia Cucu from Politehnica University of Bucharest, whose expertise has been invaluable to my research. His continuous support and feedback during the writing of my papers have played a significant role in my growth as a researcher.

I would also like to extend my gratitude to the network I have forged through IEEE. Serving as the chair of the IEEE student chapter at the University of Galway has granted me the opportunity to meet an array of incredible individuals worldwide. One particular memory that stands out is the IEEE SYP congress in Tunisia, where I had the privilege to interact with numerous talented and outstanding individuals from various corners of the globe. It was an experience that not only enriched my professional network but also led to the formation of lifelong friendships.

Vanessa Saldanha, my very first friend in Ireland, holds a special place in my heart. Your unwavering presence from the outset of my PhD journey has been an incredible source of strength and support. Thank you for everything. I'd also like to express my heartfelt gratitude to Vineet Rana for being an exceptional housemate, always ensuring I was well-fed and taken care of. To Sweta Sinha and Mohammad Saif, your boundless positivity has been a beacon of light during even the darkest times. And to the rest of our gang - Kanishka Karara, Jivesh Punjabi, and Clare Gomes - your unwavering presence has made Galway feel like home. I will always treasure our countless memories together.

I'm also thankful to my Dublin friends: Siddhant Bagul, Anusha Suvarna, Vaishakh Menon, and Sushmita Masurkar, for always providing me with a place to stay. Their efforts to cultivate a sense of home, with shared Indian cuisine, festive gatherings, and unwavering support, have truly enriched my experience in Dublin.

A heartfelt thanks also goes out to all my friends from India, from my childhood to my bachelor's journey. Your support and friendship have shaped me into the person I am today. While I can't possibly fit all your names in this acknowledgment, please know that I am immensely grateful for each and every one of you. You truly exemplify the saying, "Friends are the family we choose," and I consider myself blessed to have such amazing friends in my life.

Lastly, I want to extend a heartfelt thank you to my family. Virender Jain, my dad, has been a constant pillar of support throughout my life. His influence led me to choose computer science studies, and he has stood by me through every decision, no matter how unconventional. He has always pushed me to try things in life out of my comfort zone. I am deeply grateful for everything he has done for me since childhood. My mom, Kavita Jain, has been my rock, always there with her unwavering love and support. Her kindness and love have shaped me into the person I am today, and I aspire to spread that same warmth to others I encounter in life. To my brother, who has been my companion since childhood, I am thankful for the countless memories and the fun times we've shared, no matter the circumstances. To the rest of my family members - my uncle, my aunt, my grandma, and my little cousin Udit - your presence and support have been invaluable to me. I am deeply thankful for each and every one of you.

Quoting a phrase from Snoop Dogg, “Last but not least, I would like to thank me for believing in me, I want to thank me for doing all this hard work, I want to thank me for having no days off, I want to thank me for never quitting”. Without self-belief, none of this would have been possible. In the spirit of these sentiments, I express my gratitude to myself for the perseverance, dedication, and integrity that have guided me throughout this journey.

Abstract

Text-to-speech (TTS) and Speech-to-Text (STT) technologies have seen significant improvements in recent years with the introduction of Deep learning-based data-driven approaches, yet the application of these technologies to child speech presents unique challenges. Most current research work and solutions focus largely on adult speech compared to child speech. The main reason for this disparity can be linked to the limited availability of children's speech datasets and poor data quality that can be used in training modern speech Artificial Intelligence (AI) systems. Child speech datasets often have noisy recordings and lack diversity, resulting in limited, poor quality and less representative datasets for developing effective solutions. Child speech is also notably different from adult speech due to distinctive linguistic and phonetic characteristics, alongside variations in pitch, articulation, and pronunciation. These differences present substantial challenges in the development of effective TTS and STT systems for children. Moreover, ethical considerations and GDPR compliance necessitate careful handling of child speech data, emphasizing the need for legally compliant data collection methods. The shift to DNN and AI-based systems has improved the capacity to train on limited child speech data. Nevertheless, the availability of data remains a challenge, especially when striving to represent the linguistic and phonetic patterns of children from diverse backgrounds.

Our research focuses on several key areas: the enhancement of TTS and STT technologies for child speech in a low-resource scenario, the creation and augmentation of child speech datasets, and the integration of these technologies into practical applications such as smart toys capable of interacting with and comprehending children. We explore state-of-the-art (SOTA) methodologies, including the development and optimization of Tacotron 2 and Fastpitch models for child speech synthesis, and the application of wav2vec2, Whisper, and Conformer models for improved child speech recognition. Through the utilization of advanced data augmentation methods, it is also aimed to overcome the limitations posed by the scarcity of child speech data. Additionally, our work contributes to the broader field by developing a facial animation pipeline and creating synthetic-speaking children, addressing both technological and ethical considerations in child speech processing. The main goal of this research is to not only advance the state of child speech technologies but also to ensure their ethical and effective application in smart toys. This comprehensive study represents a significant step forward in the field of speech technology, particularly in making TTS and STT systems more accessible, representative, and effective for child users. By addressing the unique challenges associated with child speech and leveraging the latest advancements in AI and deep learning, we contribute to the development of more interactive, engaging, and supportive technological solutions in this area of research.

Contents

Acronyms

TTS: Text-To-Speech
STT: Speech-To-Text
SOTA: State-of-the-Art
DNN: Deep Neural Network
ASR: Automatic Speech Recognition
Std: Standard Deviation
El: Enterprise Ireland
DAVID: Data Center Audio/Video Intelligence on Device
DTIF: Disruptive Technologies Innovation Fund
PSOLA: Pitch Synchronous Overlap Add
HMM: Hidden Markov Model
DL: Deep Learning
SPSS: Statistical Parametric Speech Synthesis
Seq2seq: Sequence-to-Sequence
DSP: Digital Signal Processing
CNNs: Convolutional Neural Networks
RNNs: Recurrent Neural Networks
RNN-T: Recurrent Neural Network Transducer
LAS: Listen, Attend and Spell
AR: Autoregressive
NAR: Non-Autoregressive
GANS: Generative Adversarial Networks
VITS: Variational Inference with Adversarial Learning for end-to-end Text-To-Speech
NLP: Natural Language Processing
GRU: Gated Recurrent Unit
MFCCs: Mel-Frequency Cepstral Coefficients
LSTMs: Long Short-Term Memory Networks
SSL: Self-Supervised Learning
WER: Word Error Rate
CTC: Connectionist Temporal Classification
MHSA: Multi-Headed Self-Attention
GLU: Gated Linear Unit
FFNs: Feed-Forward Networks
GPU: Graphical Processing Unit
MyST: My Science Tutor

UI: User Interface
PoC: Proof of Concept
SV: Speaker Verification
CLEESE: Combinatorial Expressive Speech Engine
MOS: Mean Opinion Score
DPO: Data Protection Obligations
T-SNE: T-Distributed Stochastic Neighbour Embedding
UMAP: Uniform Manifold Approximation and Projection for Dimensionality Reduction
3D: 3-Dimensional

List of Figures

List of Tables

Chapter 1

Introduction

In the field of speech technology, two pivotal areas are Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems. These technologies, crucial in bridging human-computer interaction, face unique challenges when adapted for child speech [1], [2]. This introduction aims to elucidate the intricacies of ASR and TTS as they pertain to child speech, highlighting both the challenges and the opportunities in this subfield.

ASR technology, designed to convert spoken words into text, encounters specific challenges when dealing with child speech [3], [4]. Children's speech is inherently different from adults' due to various factors such as their vocal tracts are smaller, leading to higher pitched and less articulated speech; their language skills are in developmental stages, resulting in varied vocabulary usage and sentence structures; and their speech patterns are more dynamic and less predictable [5], [6], [7], [8], [9], [10], [11]. These distinctions require an ASR system that can accurately interpret and transcribe children's speech, overcoming the limitations of traditional models predominantly trained on adult speech.

Conversely, TTS systems, which generate spoken language from text, must be adept at producing speech that sounds natural to children. This involves not just mimicking the pitch or tone of a child’s speech, but also understanding and replicating the nuances and simplicity inherent in the way children speak and process language [12], [13]. The development of child-centric TTS systems has significant implications, especially in educational and entertainment contexts, where engaging and understandable audio content is crucial [14].

This thesis addresses a critical and often overlooked aspect of speech technology: the challenges and opportunities presented by child speech, particularly treating it as a low-resource language within the domain of ASR and TTS systems. Child speech is markedly different from adult speech, characterized by its unique phonetics, fluctuating rhythm, and dynamic tonal variations. These characteristics not only pose significant challenges for conventional ASR and TTS systems, which are predominantly trained and optimized for adult speech patterns but also highlight the lack of focused research and resources dedicated to this demographic, akin to the issues faced by low-resource languages.

The primary focus of this research is to bridge this gap by building on the existing ASR [15], [16], [17], [18] and TTS technologies [19], [20], [21] and adapting them for child speech. This involves recognizing child speech as a distinct category, requiring dedicated research attention similar to that given to low-resource languages. The lack of substantial and diverse datasets for child speech further compounds these challenges, mirroring the obstacles faced in developing technologies for languages with limited digital resources. Through this lens, the thesis explores innovative methodologies and advanced deep learning methodologies to enhance the performance of speech technologies in accurately recognizing and synthesizing child speech. By improving the capabilities of ASR and TTS

systems in handling child speech, this research aims to unlock new potentials in educational technology, interactive learning tools, and child-centric applications. The adaptation of ASR and TTS technologies for child speech is not merely a technical challenge; it represents a significant step towards making digital technologies more accessible and beneficial for younger users [22], [23]. This introduction sets the stage for a deeper exploration into the specific challenges, methodologies, and impacts of ASR and TTS technologies in the domain of child speech.

It is also interesting to note that Speech-to-Text (STT), which is closely related to ASR, represents a specific implementation of ASR technology. ASR refers to the process of converting spoken language into text using machine learning algorithms to recognize and transcribe the audio input. STT, on the other hand, is a specific application of ASR where the recognized text is displayed or used for further processing, such as in voice-controlled interfaces or transcription services. While both ASR and STT involve the conversion of speech to text, ASR is the underlying technology that powers speech recognition capabilities, whereas STT emphasizes its integration with other applications or devices. In the context of this thesis, I will use the terms ASR and STT interchangeably, as the distinction is not crucial for the specific discussion at hand.

1.1 DAVID Project

The Data Center Audio/Video Intelligence on Device (DAVID) project [24] was a collaboration between XPERI[1], Soapbox Labs[2], and the University of Galway[3] funded by Enterprise Ireland[4] (EI) under the Disruptive Technologies Innovation Fund (DTIF). Its main objective was the development of a multimodal (sound and vision) AI processing platform with low cost and low power consumption to be used for the creation of voice-enabled toys. The DAVID smart-toy platform is outlined as one of the pioneering Edge-AI platform designs that integrates advanced, low-power neural inference models for data processing directly alongside image or audio sensors. This innovative platform includes the capability for on-device text-to-speech generation. The platform is equipped with a speech-driven interface and utilizes its computer vision sensor node to recognize and interpret user interactions and facial expressions. Embedded (on-device) processing of data is currently the preferred solution across the smart toy industry to enable Artificial Intelligence in smart toys [25], [26], [27].

The project aimed to refine XPERI and SoapBox Labs' technologies for the Smart Toy market, introducing new, child-specific innovations. The University of Galway contributed by developing neural-based TTS and child speech understanding technology, customizable to different voices, and enhancing existing AI solutions for intelligent toys. The University worked towards improving several state-of-the-art technologies reliant on data-center-level AI. Figure 1 shows the image of the ‘DAVID’ teddy bear prototype.

[1] https://xperi.com/
[2] https://www.soapboxlabs.com/
[3] https://www.universityofgalway.ie/
[4] https://www.enterprise-ireland.com/en/

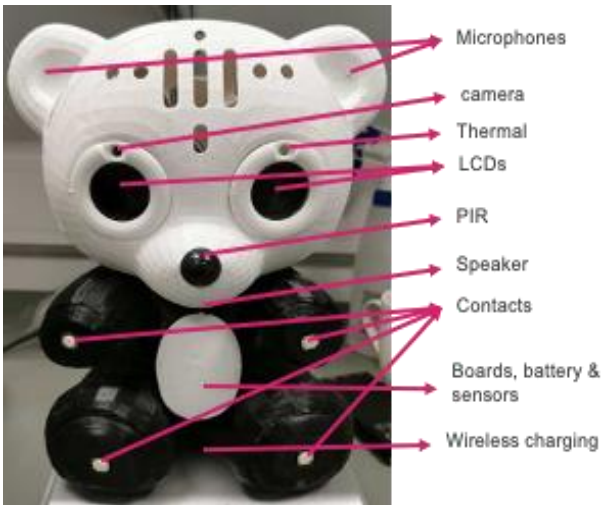


Figure 1: DAVID Smart-toy Prototype.

The Ergo AI processor, developed by Perceive, a subsidiary of Xperi, is designed to run data center-class neural networks in power-constrained environments. This makes it highly suitable for a wide range of applications in edge devices, where power efficiency is crucial. The Ergo chip by Perceive is notable for its ability to deliver high-level inferencing capabilities at ultra-low power, catering to the growing demand for smart, power-efficient devices in various sectors. It is aimed to utilize this chip as a part of the DAVID project to make the toy alive. Its efficiency in running advanced neural networks with low power consumption makes it ideal for edge devices like a smart teddy. More details about the DAVID smart toy platform and Ergo chip are present in our published work in Appendix H.

I began my academic journey as a Research Assistant at the University of Galway, primarily contributing to the development of TTS and ASR technologies with a focus on child speech, as detailed in this thesis. My involvement in various research areas, and the motivations driving these pursuits, are elaborated upon in subsequent chapters. The DAVID project, in particular, was a significant turning point. Within a year of engaging with this project, I discovered a deep-seated passion for research, which led me to transition into a full-time PhD program. This decision was largely influenced by my experiences and the insights gained during my tenure as a Research Assistant, signalling a commitment to advancing the field of speech technology. Therefore, I enrolled at the University of Galway as a PhD student with the aim to further improve speech technologies, particularly focusing on the nuances of child speech and the challenges associated with it. This pursuit was not just an extension of the DAVID project, but a deeper commitment to advancing the field of speech technology, driven by a desire to make meaningful contributions to an area rich with potential for innovation and impact.

1.2 Overview of the Main Contribution to This Thesis

The following present the core contributions of this thesis which are summarized in the below sub-sections. In the remaining chapters of this thesis, the work related to these contributions is presented. In each chapter, an introductory paragraph provides the context of the research work. Following that, the research objectives of the work are given, followed by the contributions of the presented research work. In the course of this thesis, we have significantly contributed to the academic community through the authorship of five journal papers and five conference papers. Detailed information regarding these

publications can be found in Section 1.4. For a comprehensive review and further study, the complete texts of these papers are provided in the Appendix. This thesis makes several novel scientific contributions that have been explored in various research efforts. It explores the fundamental aspects of how and why child speech differs from adult speech, providing crucial insights for understanding the unique considerations needed in speech technology development for children. The thesis also analyses existing child speech datasets, discussing the processes involved in cleaning and preprocessing them, and addresses the development of an application designed for the collection of child speech data. The thesis presents one of the first published works dedicated to generating child speech in low-resource settings, leveraging state-of-the-art TTS methodologies, and developing novel validation methodologies. Additionally, it details advancements in ASR technologies for child speech, with the results outperforming any previously obtained on the same datasets and also improves on unseen native and non-native English child speech datasets. Furthermore, the thesis focuses on the development of synthetic child speech datasets and augmentation methodologies, providing valuable insights into the distinct characteristics that differentiate adult and child speech. These datasets are also released publicly for research use. Finally, the thesis highlights additional contributions, including the organization of a special session and collaboration with industry partners, as well as the creation of innovative pipelines for animating facial landmarks and generating synthetic-speaking children.

1.2.1 CONTRIBUTION TOWARDS CHILD SPEECH DATASET CREATION

Chapter 3 of this thesis is dedicated to advancing the creation of datasets specifically tailored for child speech research. This chapter thoroughly explores the various challenges inherent in dealing with child speech. It begins by delving into the fundamental aspects of child speech, emphasizing how and why it significantly differs from adult speech. This exploration is crucial for understanding the unique considerations needed in speech technology development for children.

The chapter then transitions to an in-depth analysis of the child speech datasets that are currently available and utilized in this project. It discusses the processes involved in cleaning and preprocessing these datasets to make them conducive for training in ASR and TTS systems. This section not only outlines the technical steps but also reflects on the intricacies and nuances involved in preparing child speech data for technological applications. Additionally, the chapter addresses the development of an application designed for the collection of child speech data. This part of the chapter details the design considerations, functionalities, and the overall importance of such an application in the broader context of speech technology research. This application plays a pivotal role in enhancing the quality and quantity of child speech data, which is essential for the continued development and refinement of ASR and TTS technologies for young users.

1.2.2 CONTRIBUTION TOWARDS IMPROVING CHILD SPEECH GENERATION

Chapter 4 of this thesis presents an in-depth examination of the TTS methodologies employed within this research, focusing on state-of-the-art (SOTA) technologies like Tacotron 2 [28] and Fastpitch [21]. This chapter outlines the development of a transfer learning pipeline specifically designed for child speech synthesis in low-resource settings. Significantly, this research represents one of the first published works dedicated to generating child speech in such scenarios. In this chapter, two key publications detailing our work on Tacotron 2 [29] and Fastpitch [30] are discussed. The primary objective was

to investigate and adapt cutting-edge TTS methodologies for child speech synthesis. The research involved a comprehensive exploration of various TTS solutions, with a concerted effort to modify and optimize these systems for child-specific TTS applications.

Furthermore, the chapter delves into the novel subjective and objective validation methodologies that were developed as part of this research. These methodologies are essential for the accurate evaluation of the synthesized child speech, ensuring the quality and effectiveness of the TTS models. The validation processes not only assess the technical performance of the TTS systems but also gauge their effectiveness in realistically replicating the nuances of child speech.

1.2.3 Contribution Towards Improving Child Speech Recognition

Chapter 5 of the thesis is dedicated to detailing the advancements made in ASR technologies, tailored for child speech. This chapter provides a deep exploration of three principal ASR methodologies: wav2vec2 [17], Whisper [18], and Conformer [16]. The focus here is not only on the theoretical underpinnings of these models but also on their practical applications in the domain of child speech recognition. A significant portion of the chapter is devoted to a comprehensive series of experiments and a comparative analysis conducted between these models using a variety of child speech datasets. This resulted in the four publications as documented in references [31], [32], [33] and [34]. These experiments were meticulously designed to maximize the efficacy of training with these models, aiming to offer deep insights into their performance across different child speech datasets, including both seen and unseen datasets.

Moreover, the chapter engages in a thorough discussion of the findings from these experiments, extracting key takeaways and lessons learned. The insights gained from these analyses are instrumental in understanding how ASR technologies interact with child speech, highlighting both the strengths and limitations of current methodologies. In addition to presenting the immediate results of the research, this chapter sets a solid foundation for future work in the field. It establishes a baseline for the ongoing development and refinement of ASR technologies for child speech, ensuring that subsequent research can build upon the substantial work already accomplished.

1.2.4 Contribution Towards Child Speech Augmentation Methodologies

In Chapter 6, the focus is on the development of synthetic child speech datasets and the application of augmentation methodologies for transforming adult speech into child speech. This chapter delves deeply into the processes and techniques involved in creating these synthetic and augmented datasets, which are significant outcomes of this research. The chapter not only details the technical aspects of dataset creation but also discusses how these newly developed resources have been made accessible to the wider research community, facilitating further studies in the field.

Additionally, the chapter engages in a thorough discussion on both subjective and objective evaluations of the augmented datasets. These evaluations are crucial for assessing the quality and efficacy of the synthetic speech, ensuring that it is a viable resource for research and practical applications. Through this analysis, the chapter provides valuable insights into the distinct characteristics that differentiate adult and child speech, particularly in terms of speaker embedding features. This exploration into the nuances of adult versus

child speech adds a layer of depth to our understanding of speech characteristics across different age groups.

1.2.5 Additional Contributions

Chapter 7 highlights the additional contributions made as a part of this thesis. I have collaborated with my PhD colleagues and Xperi Engineers (DAVID project) on different research publications. We first discuss our contribution towards the special session 'Research Advances in Child Speech Technologies' which was presented at the SpeD 23 conference [1] at the University Politehnica of Bucharest, Romania, organized by my supervisor Peter Corcoran and myself. Additionally, my contribution towards the DAVID project is provided which focused on the integration of TTS and STT technologies onto the Ergo platform. This task involved collaborating with Xperi's engineering teams across Ireland and the USA, encompassing a range of expertise from engineering to linguistics and product management. I also worked towards the creation of an innovative pipeline for animating facial landmarks in sync with speech, aimed at enhancing automated dubbing with a fellow PhD Student. Lastly, I worked with other PhD students on creating a comprehensive approach for making synthetic-speaking children, combining techniques in face generation, speech synthesis, and facial animation.

1.3 List of Publications

1.3.1 Contribution Towards Improving Text-To-Speech Technologies for Children

In this section, one journal and one conference paper have been published. A copy of the published papers is attached in Appendix A and Appendix D of this thesis report.

1. R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran and H. Cucu, "A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," in IEEE Access, vol. 10, pp. 47628-47642, 2022, doi: 10.1109/ACCESS.2022.3170836.
2. R. Jain and P. Corcoran, "Improved Child Text-to-Speech Synthesis through Fastpitch-based Transfer Learning," 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2023, pp. 54-59, doi: 10.1109/SpeD59241.2023.10314899.

1.3.2 Contribution Towards Enhancing Child Speech Recognition

This section will list publications related to ASR improvement for children, which include two journal papers and two conference papers. A copy of the published paper is attached in Appendix B, Appendix C, Appendix E and Appendix F of this thesis report.

3. R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran and H. Cucu, "A WAV2VEC2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition," in IEEE Access, vol. 11, pp. 46938-46948, 2023, doi: 10.1109/ACCESS.2023.3275106.

[1] https://sped.pub.ro/

4. Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., Cucu, H. (2023) Adaptation of Whisper models to child speech recognition. Proc. INTERSPEECH 2023, 5242-5246, doi: 10.21437/Interspeech.2023-935.
5. A. Barcovschi, R. Jain and P. Corcoran, "A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition," 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2023, pp. 42-47, doi: 10.1109/SpeD59241.2023.10314867.
6. R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, 'Exploring Native and Non-Native English Child Speech Recognition with Whisper', *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3378738.

1.3.3 Contribution Towards Child Speech Data Augmentation Methodologies and Validation

This section lists publication for data augmentation and synthetic data methodologies. This led to one journal paper contribution highlighted in Appendix G of this report.

7. M. Y. Yiwere, A. Barcovschi, R. Jain, H. Cucu and P. Corcoran, "Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale.," in IEEE Access, doi: 10.1109/ACCESS.2023.3317360.

1.3.4 Other Contributions

This section will list the work done as additional contributions during my PhD program. A copy of the published paper is attached in Appendix H, Appendix I, and Appendix J of this thesis report.

8. D. Bigioi, H. Jordan, R. Jain, R. McDonnell and P. Corcoran, "Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing," in IEEE Access, vol. 10, pp. 133357-133369, 2022, doi: 10.1109/ACCESS.2022.3231137.
9. G. Cosache, F. Salgado, R. Jain, C. Rotariu, G. Sterpu and P. Corcoran, "Data Center Audio/Video Intelligence on Device (DAVID) - An Edge-AI Platform for Smart-Toys," 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2023, pp. 66-71, doi: 10.1109/SpeD59241.2023.10314915.
10. M. Ali Farooq, D. Bigioi, R. Jain, W. Yao, M. Yiwere and P. Corcoran, "Synthetic Speaking Children – Why We Need Them and How to Make Them," 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2023, pp. 36-41, doi: 10.1109/SpeD59241.2023.10314943.

1.4 Contribution Taxonomy

Due to the fact that this publication-based thesis contains collaborative effort, this section gives an outline of the primary factors that identify primary authorship. The CRediT approach has been adopted by journals in several fields to specify the contributions of individual authors. In the CRediT Taxonomy, all authors' contributions are measured as a percentage point on 14 roles. These are Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review

& editing. Despite collaborations, most of the work in this thesis is my own; hence, a more compact generalization of this taxonomy that contains the primary criteria has been selected. To be more specific:

1. Research Hypothesis/ Idea.

2. Methodology comprises validation, data creation, formal analysis, instrument selection, software development, implementation, and experiments.

3. The background includes investigation, formalization, and work done to place the research efforts in a broader context of literature in a given field; this may include some aspects of writing (literature reviews) and informs aspects of project administration and supervision, as well as ensuring that the methodology employed is typical of that used in the area of publication.

4. Manuscript preparation which includes all aspects of writing manuscript preparation including Writing – the original draft, Writing – review & editing, and Visualization except those specified in the next criteria.

This generalization has the weakness that it ignores most aspects of funding, project administration, resources or supervision but otherwise encapsulates the main points that would determine primary authorship. Such a table will be presented in each main work presented in this Thesis, attributing the contribution of each author to the aforementioned four criteria. Contribution percent is listed at a resolution of %. The authors and co-authors are listed by initials where RJ means Rishabh Jain, MY means Mariam Yiwere, DB means Dan Bigioi, PC means Peter Corcoran, AB means Andrei Barcovschi, HC means Horia Cucu, MAF means Mohammad Ali Farooq and GC means Gabriel Costache. This process is similar for any other additional co-authors in the paper. This taxonomy is provided at the start of each Appendix highlighting the author's contribution towards each of those papers.

1.5 Thesis Structure

The rest of the thesis structure is as follows.

Chapter 2 presents the evolution of speech technology, specifically TTS and ASR, tracing their development from early methods to advanced neural models.

Chapter 3 focuses on the creation and importance of child speech datasets for TTS and ASR enhancement. It discusses the scarcity of such datasets, the complexities in their collection and use, and the methodologies for data cleaning and pre-processing. It also discusses the application developed for child speech data collection.

Chapter 4 details the advancements in TTS for child speech, particularly through the development and optimization of Tacotron 2 and Fastpitch models within a transfer learning framework. Further details of these contributions are detailed in journal publications 1 [29] and journal publication 2 [30] listed in section 1.3.

Chapter 5 explores enhancing ASR technologies for child speech, using models like wav2vec2 for self-supervised learning and comparing their effectiveness with other models such as Whisper and Conformer models. Further details of these contributions are detailed

in journal publications 3 [31], 6 [34] and conference publications 4 [32], 5 [33] listed in section 1.3.

Chapter 6 addresses the generation of synthetic child speech datasets, showcasing how these datasets, created using techniques like adult-to-child speech augmentation, aid in finetuning ASR models and contribute to child speech research. Further details of these contributions are detailed in conference publications 2 [30] and journal publication 7 [35] listed in section 1.3.

Chapter 7 outlines additional contributions to the DAVID project and related research, including collaborations on integrating TTS and ASR technologies for smart toys, developing audio features for facial animation, and highlighting significant advancements in creating synthetic-speaking children. It highlights the research work done in collaboration with the industry partner, colleagues and other PhD students [24], [36], [37].

Chapter 8 outlines the main conclusions and future work based on the work contained in this thesis.

Chapter 2

Introduction to Speech Technology

This chapter provides a comprehensive overview of the evolution and fundamental workings of Text-to-Speech (TTS) and Speech-to-Text (STT) technologies. Initially, it delves into the historical progression of TTS, tracing its development from early synthesis methods to contemporary advanced models. Similarly, the evolution of STT is explored, highlighting the significant milestones from its inception to the sophisticated systems used today. The chapter then shifts focus to the core principles and mechanisms underlying TTS and STT technologies, offering a detailed understanding of how these systems convert text to speech and vice versa. Special attention is given to the workings of specific TTS models, namely Tacotron 2 [28] and Fastpitch [21], which are utilized in this work. These models represent the forefront of TTS technology, and their operational intricacies are crucial for comprehending their application in speech synthesis in this thesis. The chapter also examines STT models used in this research, including wav2vec2 [17], Whisper [18], and Conformer [16]. By dissecting the functionalities of these state-of-the-art STT models, the chapter provides insight into the latest developments in the field of ASR.

2.1 Synergy between Speech-To-Text and Text-To-Speech

The synergy between STT and TTS systems lies in their complementary nature, where improvements in one can directly influence advancements in the other [1], [38], [39], [40]. These technologies are intrinsically linked through their shared goal of facilitating natural human-computer interaction and their underlying technologies. When considering child speech, the synergy between ASR and TTS becomes particularly crucial due to the unique challenges posed by limited data availability and the distinct characteristics of child speech (will be discussed in detail in Chapter 3). The **TTS systems** can synthesize childlike speech to create diverse datasets. This is especially useful where collecting large volumes of natural child speech is challenging due to ethical and logistical considerations [41]. Generated speech data can encompass various accents, dialects, and speech patterns, contributing to a more robust training set for child-specific ASR systems. On the other hand, **ASR systems** can be used to transcribe child speech in educational settings, speech therapy, or interactive learning applications. Transcriptions generated by ASR provide valuable data for linguistic research and the development of child-specific language models.

Upon initiating our research on child speech, it became apparent that the research domain was experiencing a scarcity of datasets comprising child speech (discussed in more detail in Chapter 3), which are critical for various speech recognition and synthesis projects. Originally, our research endeavoured to engineer a TTS system, specifically for integration with the DAVID platform which will be able to provide controllable child speech synthesis [42], [43]. Confronted with the dual challenges of dataset paucity and the prevalence of unannotated child speech collections, our investigative trajectory shifted toward an

exhaustive examination of ASR technology with a particular focus on child speech patterns.

As our exploration into ASR for child speech deepened [3], it became increasingly evident that there exists considerable potential for advancement in this specialized ASR domain. Such enhancements have the dual benefit of enriching the transcription of child speech corpora and, by extension, augmenting the performance and accuracy of TTS systems [38], [44]. This symbiotic enhancement is pivotal, as it promises to yield more comprehensive and representative child speech datasets, thereby significantly contributing to the broader research ecosystem within this specialized field.

The interplay between ASR and TTS addresses the critical issue of data scarcity in child speech research. By using TTS to augment existing datasets and ASR to generate new transcriptions, researchers and developers can overcome the hurdle of limited child speech data. This synergy enables the development of more effective, accurate, and inclusive speech technologies tailored for children, fostering advancements in educational technology, speech therapy, and child-centric applications [45], [46]. For this reason, we decided to tackle the issues of both TTS and ASR linked with child speech to improve overall child speech understanding using AI-based development.

2.2 Evolution of Text-To-Speech

The historical development of speech synthesis has seen a shift from early attempts using parametric synthesis methods, such as Wolfgang von Kempelen's [47] machine in 1971, to the introduction of Klatt's serial/parallel formant synthesizer [48] in 1980. The DECtalk text-to-speech system [49] in 1990 improved speech quality with the Pitch Synchronous OverLap Add (PSOLA) algorithm [50]. However, challenges persisted, leading to the exploration of advanced models like the Hidden Markov Model (HMM)-based [51], [52] and Deep Learning (DL)-based [13] synthesis methods.

Traditional speech synthesis involves two main approaches: concatenative TTS and parametric TTS [53]. Concatenative synthesis [54] concatenates pre-recorded speech units to form a continuous stream, with schemes like LPC-based and PSOLA-based methods. The former preserves speech information but lacks natural flow, while the latter addresses prosody control issues. Parametric synthesis [53] leverages digital signal processing to simulate the vocal process, offering various methods like vocal organ and formant parametric synthesis, HMM-based [55], and DNN-based synthesis [12]. The Statistical Parametric Speech Synthesis (SPSS) employs three modules: text analysis, parameter prediction, and speech synthesis. It utilizes linguistic features to enhance naturalness and quality, demonstrating significant improvements in experimental results.

Deep Learning (DL) has revolutionized the speech synthesis [12], departing from HMM-based methods. Deep BLSTM-based models, utilizing Bidirectional Long Short-Term Memory networks, and sequence-to-sequence (seq2seq) networks have shown remarkable efficiency in mapping linguistic features to acoustic features [12]. The end-to-end speech Synthesis methods integrate text analysis, acoustic modelling, and speech synthesis into a unified framework, eliminating the need for extensive domain expertise and minimizing errors. Neural TTS emerged as a paradigm shift, employing neural networks as the backbone for speech synthesis. Early models like WaveNet [56] and DeepVoice 1/2 [57], [58] integrated neural networks into SPSS components. End-to-end models like Tacotron 1/2 [19], [28], Deep Voice 3[59], FastSpeech 1/2 [60], [61], and EATS [38] streamlined

text analysis and directly generated waveforms from text. Neural network-based synthesis offers superior voice quality, intelligibility, and naturalness while reducing human pre-processing and feature development requirements. This was a summary of the evolution of TTS, however, this [13] interesting article can be referred to for more details.

The journey of speech synthesis has progressed from early parametric methods to sophisticated deep-learning models, achieving remarkable strides in naturalness, intelligibility, and overall speech quality. Recent models like Tacotron 2 and Fastpitch showcase state-of-the-art advancements, emphasizing end-to-end synthesis and efficient parallel computation. For this reason, they were also considered as primary TTS models to be used for child speech synthesis in this research work. This will be discussed in more detail in section 2.4.

2.3 Evolution of Speech to Text

In 1952, Bell Laboratories introduced the "Audrey" system [62], marking an early effort in recognizing spoken numbers. The 1960s saw progress fueled by digital signal processing (DSP) and pattern recognition algorithms [62]. The 1970s brought about the influential "Hidden Markov Model" (HMM) [63], as outlined by Lawrence R. Rabiner in 1989, setting the stage for modern ASR systems. Advancements in neural networks [64], [65] during the 1990s and 2000s laid the foundation for large vocabulary systems, while the 2010s witnessed breakthroughs with deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Big data and cloud computing further empowered ASR systems, leading to consumer applications like virtual assistants. The emergence of models like Listen, Attend and Spell (LAS) from Carnegie Mellon University [66] and Deep Speech 2 [67] from Baidu in 2015 showcased the power of encoder-decoder architectures and deep learning in achieving state-of-the-art performance.

RNN-Transducer (RNN-T) [68], [69], introduced transformer and transducer architectures, respectively, achieving top-tier performance. The year 2020 brought wav2vec2 [17] by Meta AI Research [1], utilizing self-supervised learning with a CNN architecture. Conformer's [16], [70] marked a new era with a conformer architecture and streaming attention mechanism for real-time, efficient speech recognition. In 2022, OpenAI's[2] release of Whisper [18] added a noteworthy contribution to ASR, providing open-source models ranging from tiny to large, further diversifying the landscape of speech-to-text technology. Looking ahead, recent models such as wav2vec 2, Whisper and Conformer are pushing the boundaries of ASR. These models represent the ongoing innovation and diversity in the field of AI-based speech recognition. Due to their outstanding performance with adult speech datasets, they are also used for improving the child speech ASR in this thesis.

2.4 Neural Text-To-Speech Technologies

The basic architecture of a Neural TTS system (see Figure 2) involves three main components:

[1] https://ai.meta.com/research/
[2] https://openai.com/

Figure 2: Key Components of neural Text-To-Speech.

1) Text Encoding module: This component converts the input text into a numerical representation that can be processed by the deep neural network. Common techniques used in text encoding include word embeddings, character embeddings, and RNNs.
2) Acoustic model: This component generates the acoustic features of speech, such as pitch, tone, and spectral envelope, based on the encoded text. The acoustic model typically employs a deep neural network, such as a CNN or an RNN, to predict the acoustic features from the text encoding. The features provide a time-frequency representation of the audio waveform, like a spectrogram.
3) Vocoder: This component converts the acoustic features produced by the acoustic model into a digital audio signal that can be played back as speech. The vocoder typically uses a waveform generator network, such as a Griffin-Lim algorithm [71] or a WaveNet [56], which takes in the spectrogram as input and produces a waveform that closely matches the original speech signal.

2.4.1 Selection and Evaluation of the TTS model

In our comprehensive exploration of TTS technologies, we delved into the latest advancements and research in the field to identify the most effective and efficient solutions [13]. Originally, the goal was to identify TTS models that were compact, compatible with PyTorch, easy to implement, and suitable for deployment on edge devices. Additionally, there was a focus on developing methodologies to adapt these models for effective use with child speech.

2.4.1.1 Acoustic Model Selection (Mel Spectrogram Synthesizer)

The emergence of the Tacotron [19] represented a paradigm shift in the field of speech synthesis, significantly elevating the quality of synthesized speech. Despite the advent of newer methodologies boasting enhanced efficiency and smaller model architectures, Tacotron continues to be a serving as a benchmark for quality assessment in comparison to emerging techniques. In the course of our research, we conducted an extensive exploration of various cutting-edge TTS models including Tacotron. We performed an in-depth analysis of various state-of-the-art systems and methodologies mentioned in Table 1. From our preliminary evaluation of TTS models, we divided the models based on the following categories: Number of parameters, Model Type (Autoregressive/Non-autoregressive), potential to work on Ergo chip, Speech quality and Pytorch implementation availability. The utilization of the Pytorch framework was imperative for our objectives, as it was compatible with Xperi's proprietary training framework for porting models over to edge devices. Table 1 presents a detailed summary of this comprehensive exploration.

Table 1: Initial Exploration of TTS Models [72]

TTS Models	Number of Parameters	Type of Model	Potential to Work on ERGO	Speech Quality Ranking
Tacotron 2 [28]	29 M	AR	Likely	2
Deep Voice 3 [59]	7 M	AR	Likely	7
TransformerTTS [73]	24 M	AR	Likely	1
Flowtron [74]	61 M	AR	Unlikely	8
Paranet [75]	17 M	NAR	Very Likely	6

FastSpeech [60]	23 M	NAR	Very Likely	5
FlowTTS [76]	NA	NAR	Unknown	NA
FastSpeech 2 [61]	27 M	NAR	Very Likely	3
Glow-TTS [77]	28.6 M	NAR	Very Likely	4

M=Millions, AR=Autoregressive, NAR=Non-Autoregressive.

Table 1 presents a comprehensive exploration of various TTS models, and the ranking was carefully determined based on multiple factors. Firstly, we ensured that only models with publicly accessible implementations were included in the table. The availability of public implementations allowed for transparency and reproducibility, enabling us to evaluate each model's performance and characteristics. We also engaged with the TTS community, including developers, researchers, and users, to gather their insights, experiences, and subjective evaluations of the listed models. Ranking these models accurately is challenging because they vary in training data, duration, and computing resources. Therefore, we only assessed the audio quality of sample outputs from the models' GitHub repositories. We listened to and analysed the synthesized speech, taking into account factors such as naturalness, clarity, and overall intelligibility. Community feedback also played a pivotal role in our ranking process such as going through code reviews, issues, and discussions. Table 1 was developed during the initial phase of our research, and we acknowledge that numerous updates and variations have likely been introduced since then. The ranking provided in Table 1 is based on evaluations of sample outputs, literature reviews, and community feedback, and it is open to revision. We encourage readers to consider Table 1 as a starting point and to refer to the latest advancements and updates.

These models predominantly supported single-speaker synthesis and therefore we focused on training these TTS models using the single-speaker LJ Speech dataset [78]. This initial training phase was crucial in establishing a baseline understanding of the models' capabilities and limitations. We selected the models which looked promising for further training with child speech. Despite the theoretical promise of these models, the experimental results did not meet our expectations when applied to TTS, particularly with child speech. One key issue is that these models are heavily dependent on large, diverse training datasets to capture the full range of phonetic and prosodic features present in natural speech. Child speech, with its unique characteristics and variations, presents an even greater challenge, as the models must learn from a significantly smaller and less varied dataset compared to adult speech. Common challenges included the lack of naturalness in speech synthesis, difficulties in capturing the emotional subtleties of speech, and the inability to effectively replicate unique speech patterns, such as those found in child speech. Additionally, many systems struggled with the accurate pronunciation of uncommon words and names, and the adaptability to different languages and dialects when used with child speech.

During our research, we also came across variations in TTS models with multispeaker capabilities such as Deep Voice 2 [57]. The development of DeepVoice 2 marked a significant advancement, integrating speaker verification models to facilitate multispeaker TTS synthesis [20], [79], [80], [81], broadening the scope of TTS applications. The methodology proposed by DeepVoice 2 can be incorporated with other TTS models such as Tacotron 2, WaveNet, and more recent innovations like DeepVoice 3 [59] or VITS [82]. These models incorporate variational autoencoders and attention mechanisms, among other techniques, to learn the subtleties of multiple speakers' voices within a single framework. This discovery was pivotal, as it aligned more closely with our objective of enabling TTS

for child speech. The introduction of these multispeaker-capable models [20], [80], [83] significantly expanded our research scope and prompted a shift in our approach.

Consequently, we undertook a comprehensive review of TTS models, taking into account various factors as described in Table 2. These included the number of parameters, the type of model, speech quality, multi-speaker compatibility, and real-time processing capability, along with each model's specific advantages and disadvantages. The intricacies involved in tailoring these models to accurately represent child speech became a more pressing and fundamental goal, taking precedence over their immediate implementation on Edge devices. Therefore, It was decided to adapt the TTS model for child speech synthesis as the primary focus. Consequently, this led to an adjustment in focus, with less emphasis being placed on Ergo compatibility for the time being. This decision was also influenced by the parallel efforts of Xperi engineers, who were concurrently working on optimizing TTS models for Edge devices.

Table 2: Review of TTS Models

Model	Para-meters	Model Type	Multi-speaker	Real-Time	Advantages	Disadvantages
Tacotron [19]	~28.2M	AR	Yes	Slow	Pioneer in end-to-end TTS	Slower, less robust
Tacotron 2 [28]	~29M	AR	Yes	Slow	Improved over Tacotron	Still slower inference
Deep Voice 2 [57]	~30M	AR	Yes	Medium	Good for multispeaker TTS	Lower quality than newer models
Deep Voice 3 [59]	~20M	AR	Yes	Medium	Better voice quality than Deepvoice2	Outperformed by newer models
Transformer TTS [73]	~44M	AR	No	Slow	High-quality voice	Very resource-intensive
Flowtron [74]	~61M	AR	Yes	Slow	Flexible voice style	Very large model size
Paranet [75]	~17M	NAR	No	Fast	Fast, lightweight	Lower naturalness
Fastspeech [60]	~23 M	NAR	No	Fast	Faster than autoregressive	Compromised audio quality
Fastspeech 2 [61]	~23 M	NAR	No	Fast	Improved quality over Fastspeech	Complex training pipeline
Flow-TTS [76]	NA	NAR	No	Medium	Quality similar to Flowtron	Large, complex model
Glow-TTS [77]	~29 M	Flow-based	Yes	Fast	Good balance of speed and quality	Requires finetuning
SpeedySpeech [84]	~30 M	NAR	Yes	Fast	Very fast, efficient	Compromises on quality
VITS [82]	~30 M	AR	Yes	Medium	State-of-the-art quality	Very resource-intensive
Fastpitch [21]	~30 M	NAR	Yes	Fast	Fast, high-quality	Requires high-quality training data

M= Millions, AR= Autoregressive, NAR= Non-Autoregressive.

After thorough evaluation and experimentation, we found that Tacotron 2 [28], particularly when adapted for child speech, outperformed other models in several key areas. Tacotron 2's neural network architecture demonstrated superior ability in generating more natural-sounding and expressive speech. Its strength lies in its end-to-end generation capability, which simplifies the speech synthesis process and enhances the overall quality of the output. Specifically, when applied to child speech (more in Chapter 4), Tacotron 2 effectively captured the unique tonal and articulatory characteristics that are typically challenging for conventional TTS systems. The model also benefits from its extensive open-source codebase with an active research community, continuously working on innovations and variations of the model such as the multispeaker adaptation of Tacotron 2. Given its robust performance and adaptability, Tacotron 2 was chosen as the primary technology for our research into child speech synthesis and it also provides a baseline for future research.

Further, we also used the Fastpitch [21] model for our main experiments. It was after we started working with Fastspeech 2 [61] that we came across Fastpitch. The introduction of Fastpitch effectively addressed a notable limitation in Fastspeech 2: its inability to support multiple speakers. Fastpitch distinguished itself primarily through its exceptional pitch control capabilities, a feature crucial for accurately capturing the unique pitch variations of child speech. Moreover, FastPitch excelled in speech synthesis speed and efficiency, addressing one of the key limitations of Tacotron 2. The addition of features such as pitch control, speed, expressiveness, efficiency and self-attention mechanism solidified Fastpitch as our model of choice for child speech synthesis, significantly elevating the quality and realism of our synthesized speech outputs. This will be highlighted in more detail in Chapter 4.

2.4.1.2 Vocoder Model Selection (Waveform Generation)

Our exploration of vocoders did not involve extensive training. When evaluating their performance with child speech spectrograms, we observed consistent behaviour across most models. Our primary goal was to identify a universal vocoder [85], [86], [87], [88], capable of effectively handling both adult and child speech. Therefore, a specific criterion for the selection of the optimal vocoder was employed, taking into account various factors like Parameters, Model Type, Universal Vocoding Capabilities, and Real-time Capabilities, as well as their respective advantages and disadvantages. A detailed overview of this selection process is presented in Table 3.

Table 3: Review of Vocoder Models

Vocoder	Parameters	Model Type	Real-Time	Universal Vocoder	Advantages	Disadvantages
WaveNet [56]	~4.6 M	AR	No	No	High-quality speech synthesis	Very slow inference, computationally intensive
WaveRNN [85]	~4 M	RNN	Yes	Yes	Good balance of quality and efficiency	Less natural than WaveNet
Parallel WaveNet [89]	~4.6 M	NAR	Yes	No	Faster than WaveNet, high quality	Complex training process
WaveGlow [90]	~87.9 M	Flow	Yes	Yes	High-quality speech, fast inference	Large model size, intensive training

MelGAN [91]	~4.26 M	GAN	Yes	Yes	Fast, lightweight, suitable for real-time applications	Quality lower than WaveNet
LPCNet [92]	~1 M	RNN	Yes	Yes	Efficient for real-time applications, low computational load	Quality is not on par with larger models
Griffin-Lim [71]	NA	Algorithm-based	Yes	Yes	Simple, no training required	Lower quality compared to neural vocoders
HiFi-GAN [93]	~13.9 M	GAN	Yes	Yes	High fidelity, efficient for real-time applications	Requires careful tuning to avoid artifacts

M= Millions, AR= Autoregressive, NAR= Non-Autoregressive, RNN= Recurrent Neural Network, GAN= Generative Adversarial Networks

WaveNet [56], with its superior audio quality, sets a high benchmark in the field but is limited by its slow processing speed, which hinders its suitability for real-time applications. Parallel WaveNet improves upon this with enhanced processing efficiency while maintaining high-quality output. WaveRNN [85] and LPCNet [92] offer a balanced approach, with the former providing a compromise between audio quality and processing efficiency, and the latter leaning towards real-time application suitability due to its lower computational demands. On the efficiency frontier, MelGAN [91] and HiFi-GAN [93] excel with incredibly fast inference and smaller model sizes, though they slightly compromise on audio quality compared to their more computationally intensive counterparts. These characteristics make them ideal for real-time and resource-constrained environments.

WaveGlow [90] and WaveRNN [85] emerged as the optimal choices of vocoder models, a decision that was influenced by a combination of key factors. Primarily, these models demonstrated superior performance in preserving the distinct characteristics of child speech. The evaluation process involved passing child speech spectrograms through the vocoders and subjectively assessing the quality of the audio waveforms generated. WaveGlow offered the most natural-sounding voice among the tested vocoders. Its swift inference speed was commendable, but its extensive memory requirement, marked by a large parameter size, presented challenges for integration with memory-constrained devices like ERGO. WaveRNN also emerged as a favourable choice for a vocoder due to its balanced attributes. It adeptly combines high-quality audio output with computational efficiency, making it ideal for a range of applications, including those with limited processing power. WaveGlow's natural-sounding voice synthesis and WaveRNN's efficient yet quality-oriented performance, coupled with their ease of implementation and universal vocoding capabilities made them an ideal choice for our research. The chosen vocoders were also favoured due to their compatibility and ease of integration with acoustic models such as Tacotron 2 and Fastpitch. This aspect of the research, focusing on the synergy between vocoders and acoustic models will be discussed in greater detail in the subsequent sections of this chapter.

2.4.2 TACOTRON 2 WITH WAVERNN

Tacotron 2 [28] is a neural network architecture designed for TTS synthesis, employing a recurrent sequence-to-sequence (seq-to-seq) feature prediction approach. This model transforms input text characters into embedded sequences through natural language processing (NLP) tools [94], utilizing a recurrent sequence-to-sequence feature to predict Mel spectrogram sequences. The Mel spectrum is employed as a visual representation of audio data in the time-frequency domain, where the bins correspond to pitch classes. The model generates a time-domain waveform from the Mel spectrum using a modified version of the WaveNet [56] architecture. This involves performing an inverse Fourier Transform, converting data from the time-frequency domain to the time-power domain. The use of two distinct acoustic representations facilitates separate training of these components. The architecture of Tacotron 2 can be seen in Figure 3. In Tacotron, mel-spectrograms are computed through a short-time Fourier transform with a 50ms frame size and a Hann window function, which smoothens frequencies for a more visually coherent waveform. This pre-processing aids network analysis and prediction while optimizing processing efficiency. The two acoustic representations are then fed into a neural network comprising an encoder and a decoder. The encoder translates character sequences into a hidden feature representation, and the decoder utilizes this representation to predict a spectrogram. The decoder functions as an autoregressive recurrent neural network, predicting a Mel spectrogram frame by frame from the encoded input sequence. Tacotron 2 also employs an attention mechanism to allow the decoder to focus on different parts of the input sequence as it generates the output, improving the model's ability to capture long-range dependencies.

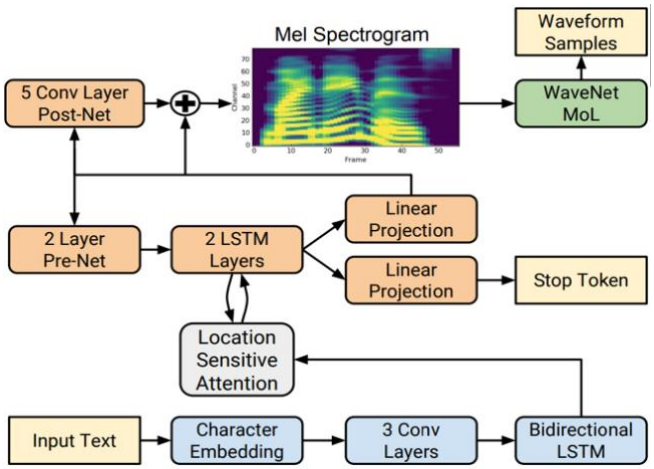


Figure 3: Architecture of Tacotron 2 [28].

We use the WaveRNN vocoder [85], which is an improvement over the WaveNet architecture employed with Tacotron 2. WaveRNN uses a Gated Recurrent Unit (GRU) in comparison to convolutions used in WaveNet. The input melspectrograms and their corresponding waveforms are segmented at each timestamp. The resulting output synthesizes a vocoder that effectively mimics the nuances of human speech. This comprehensive process ensures the generation of a high-quality and natural-sounding speech synthesis. The investigation of vocoders specifically designed for child TTS applications is a relatively unexplored area in the field of research. Since, the WaveRNN model (as cited in [85], [95]) has been acknowledged for its versatility and effectiveness as a universal vocoder, WaveRNN trained on the LibriSpeech dataset [96] was used to be utilized as the universal vocoder for synthesizing child voices with Tacotron 2. In this work, we use a modified version of Tacotron 2 which incorporates speaker embeddings. These speaker embeddings allow the model to perform multispeaker capabilities. Each speaker is

associated with a unique embedding vector, which is then incorporated into the training process to allow the model to learn speaker-specific characteristics. This will be covered in more detail in Chapter 4.

2.4.3 FASTPITCH WITH WAVEGLOW

Fastpitch [21] is a fully parallel TTS model that extends the architecture of Fastspeech [60], introducing a conditioning mechanism on fundamental frequency contours. In the inference phase, Fastpitch predicts pitch contours, allowing for dynamic alterations in the generated speech. This not only enhances expressiveness but also ensures a more cohesive match with the semantic content of the utterance. The model's simplicity, efficiency, and potential for multispeaker scenarios also add to its advantages. The underlying architecture of Fastpitch is built upon a fully parallel transformer, distinguishing it from Tacotron 2. This design choice significantly improves the real-time factor, making it more efficient in synthesizing mel spectrograms for a typical utterance. Fastpitch's architecture as detailed in Figure 4, includes two feed-forward transformer stacks that predict the duration and average pitch of every character. It emphasizes the model's ability to predict and use pitch in a low resolution, allowing for easy pitch adjustment and practical applications. Furthermore, fastpitch integrates an unsupervised speech-text aligner [97], contributing to its robust and versatile performance. Another standout feature of Fastpitch is its duration predictor which accurately predicts the duration of each phoneme in the input text, allowing the model to control the timing and rhythm of the speech more effectively. The duration predictor is a significant improvement over Tacotron 2 [28], which doesn't inherently control phoneme duration as effectively.

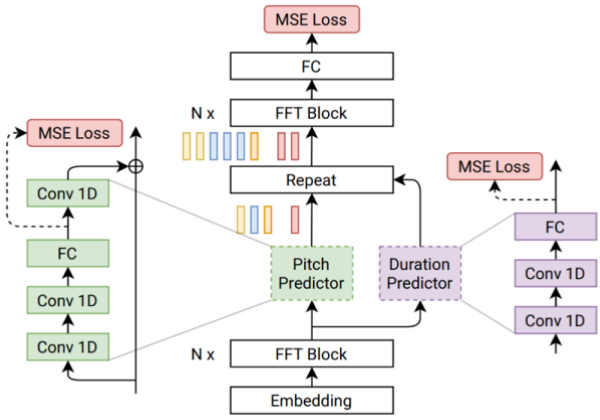


Figure 4: Architecture of Fastpitch [21].

WaveGlow [90], which is a SOTA vocoder model is used in this work to generate high-quality and natural-sounding speech waveforms from the Fastpitch output. WaveGlow belongs to the category of normalizing flow models, which are generative models that learn a one-to-one mapping from a simple distribution. WaveGlow operates by taking a spectrogram representation of the speech as input and generating the corresponding waveform. The model employs an invertible neural network to transform the spectrogram into a latent space representation and then uses a series of invertible coupling layers to map this latent representation back into the waveform domain. WaveGlow can also be conditioned on additional information, such as speaker embeddings or linguistic features, to allow for more control and customization of the generated speech. The synergy between FastPitch and Waveglow not only enhances the expressiveness and semantic alignment of the synthesized speech but also underscores the model's ability to cater to multispeaker scenarios. This collaborative approach signifies a step forward in the domain of text-to-

speech synthesis, offering both efficiency and high-quality output in a parallelized and expressive manner.

2.5 Neural Speech-To-Text Technologies

The working of neural STT involves several stages (Figure 5), which are explained below:

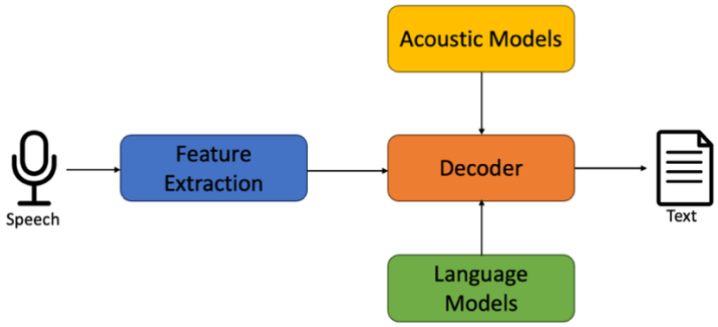


Figure 5: Key components of neural Speech-To-Text.

1. Feature extraction of input audio: The raw audio signal is converted into a feature representation that captures relevant information for speech recognition. Commonly used features include Mel-frequency cepstral coefficients (MFCCs) or spectrograms, which represent the frequency content of the audio signal over time.
2. Neural network architecture: Modern Speech-to-Text systems often use deep neural networks, such as RNNs, long short-term memory networks (LSTMs), or transformer architectures. Recurrent and LSTM networks are particularly useful for capturing temporal dependencies in speech, while transformer architectures have gained popularity for their ability to model long-range dependencies effectively.
3. Training: The neural network is trained using a dataset of paired audio and corresponding transcriptions. During training, the network learns to map the audio features to the corresponding textual representation.
4. Decoding: Once trained, the model is used for decoding. Given a new audio input, the model predicts the most likely sequence of words that corresponds to the spoken content. The highest-scoring final decoded output is elected as the final transcription.
5. Language models: Decoding often involves using a language model, which understands the rules and structure of the language. It helps in accurately constructing sentences and phrases, ensuring that the transcription makes sense in the chosen language.

It's worth noting that the quality of the STT system depends on factors such as the size and diversity of the training data, the architecture of the neural network, and the effectiveness of the decoding process.

2.5.1 Selection and Evaluation of the ASR Model

In selecting ASR models [3], [4] for this thesis, a comprehensive review was not undertaken as we did with the TTS selection. The rationale behind the chosen ASR models will become more apparent throughout this chapter. Initial tests were conducted with state-of-the-art ASR models, including Kaldi [98], DeepSpeech 1&2 [67], [99], Espresso [100], SpeechNet [101], QuartzNet [102], and Jasper [103], specifically assessing their performance on child speech. These models were selected due to their popularity with adult speech and the availability of a large community supporting it. The preliminary evaluations revealed that most models did not yield satisfactory results with child speech. We also attempted to finetune several of these models using the MyST_train child speech dataset

(more in Table 8), followed by evaluating their performance on a subsequent test set. This approach began to yield encouraging outcomes, showing a marked improvement in model performance. However, these results, while promising, did not yet reach the level of state-of-the-art achievements. It became clear that additional research and training are needed for these models to effectively adapt to and recognize child speech patterns. Consequently, we focused on using state-of-the-art (SOTA) models that are open-source and reproducible, aligning with current trends in the field.

During the Interspeech 2021 ASR Challenge, 'Shared Task on Automatic Speech Recognition for Non-Native Children's Speech' [104], we observed a significant trend in the application of the wav2vec2 [17] for child speech recognition. wav2vec2 aims to utilize Self-Supervised Learning (SSL), to learn from large amounts of unlabelled data, a crucial factor given the scarcity of labelled child speech data. The pretraining module of the wav2vec2 doesn't require transcription for audio files and can be trained with just speech data. Therefore, we intended to utilize this model to train on the unlabelled MyST dataset. It was also observed in the original wav2vec2 paper [17] that these pretraining models can be finetuned on data as small as 10 minutes of adult speech and achieve SOTA results. Therefore, it was also intended to see how this model will behave with child speech finetuning and if a low amount of child speech data be used to achieve SOTA performance. Therefore, extensive experiments were conducted with various groupings to understand the impact of different pretraining and finetuning combinations on the model's performance. We developed a comprehensive experimental framework that focuses on pretraining and finetuning, using wav2vec2 as the base model. This framework facilitates in-depth analysis and optimization of wav2vec2 for child speech. It also serves as a benchmark for comparative studies for our subsequent experiments using other ASR models. More on this will be discussed in Chapter 5.

At the time of working with wav2vec2, we also came across Whisper [18] by OpenAI. The authors of Whisper [18] successfully closed the gap in weakly supervised speech recognition by using abundant labelled audio data. They extended weakly supervised pre-training beyond English-only speech recognition to include multiple languages and multitask learning, achieving excellent performance on various multilingual adult speech datasets. Given Whisper's extensive training data, which is ten times more than wav2vec2 (680k vs. 60k) and includes many multilingual and low-resource languages, we aimed to assess how well this multilingual data can be used for child speech recognition through finetuning. The model's design, which supports multitask learning, provides advantages when dealing with the complexities of child speech, which often includes a mix of linguistic and paralinguistic elements that are challenging to capture with conventional ASR models.

We also extended the Whisper methodology to encompass additional datasets featuring non-native English child speech with diverse accents. This decision was driven by the recognition that child speech datasets can exhibit substantial variability based on the native languages of the speakers. We aimed to investigate how the model performs when exposed to a wider range of accented child speech data. Our objective was to overcome the challenges ASR systems face in accurately transcribing non-native English child speech, a notably under-researched area, especially given the limited availability of training data for these languages.

Finally, it was also decided to add a Conformer-Transducer ASR [16] in the experimental pipeline. Conformer models use their hybrid architecture, combining CNNs and

Transformers, allowing them to effectively process both local and global features of speech, an important requirement for recognizing the varied and unpredictable speech patterns of children. Conformer-transducer ASRs can also achieve competitive performance with less training data compared to Whisper, which relies on a substantial amount of labelled audio data. This can be advantageous in scenarios where collecting labelled data is challenging or expensive such as child speech. Conformers also simplify the ASR pipeline, reducing complexity and computational requirements compared to the two-stage approaches of wav2vec2. The smaller model size of Conformer Transducer models, compared to Whisper and wav2vec2, offers a distinct advantage as it reduces complexity for deployment on edge devices (like the DAVID smart toy platform). Additionally, Conformer models demonstrate robustness in handling diverse speech patterns, accents, and background noises, which are common in child speech environments. These models will be discussed in more detail in the subsequent sections.

Evaluation Metric:

For all our ASR experiments, we used Word Error Rate (WER) as the primary metric for comparisons of the results. It is the most common metric used to evaluate the performance of speech recognition systems. It measures the accuracy of transcribed speech by comparing the machine-generated transcript to the correct, human-generated transcript. WER is calculated based on the number of errors made, which are categorized into three types:

- Substitutions: When a word in the transcribed text is incorrect but is present in the speech.
- Deletions: When a word is omitted in the transcribed text but is present in the speech.
- Insertions: When an extra word is added in the transcribed text that was not present in the speech.

WER is calculated as the sum of the number of errors divided by the total number of words in the correct transcript.

$$\textbf{WER} = \frac{\text{Number of (Substitutions+ Deletions+ Insertions)}}{\text{(Total Number of Words in the Correct Transcript)}}$$

2.5.2 WAV2VEC2

The wav2vec2 model [17] can extract speech representations from raw audio files in a self-supervised learning framework, tailored for subsequent downstream ASR tasks. Notably, wav2vec2 demonstrates state-of-the-art results when trained on extensive unlabelled speech data, subsequently finetuning on labelled data, even in scenarios with minimal labelled data, as short as 10 minutes. This adaptability is particularly advantageous for tasks where acquiring accurately labelled data is challenging, such as child speech.

The model's training consists of a two-step process, as illustrated in Figure 6. The initial phase involves pretraining, wherein the model is trained on a substantial amount of unlabelled data. The subsequent step entails finetuning on labelled data utilizing the Connectionist Temporal Classification (CTC) loss [105], [106] for downstream ASR tasks. Leveraging this two-step training approach allows the model to learn speech representations in a self-supervised manner during pretraining, enabling effective training with large quantities of unlabelled speech data. This resolves challenges related to the scarcity of labelled child speech data, as the pretraining model can be trained on a combination of unlabelled child speech data and abundant adult speech data.

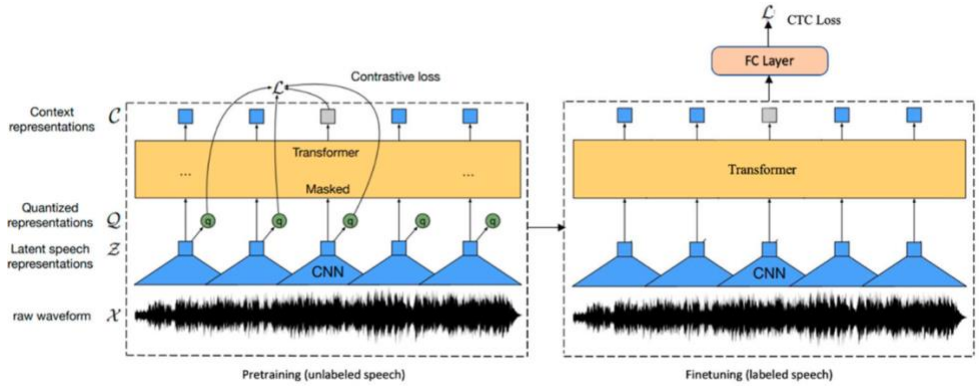


Figure 6: Pretraining and finetuning steps in the wav2vec2 architecture [17].

In the pretraining phase of wav2vec2, three key components are involved: a feature encoder, a context network, and a quantization module. The feature encoder, employing a series of 1D convolutional layers, processes the raw audio waveform, producing a sequence of feature vectors that represent the input waveform. The context network, a transformer-based encoder, further processes these feature vectors using self-attention mechanisms, facilitating the capture of long-range dependencies in the input data. The quantization module employs a codebook of fixed vectors and a Gumbel softmax function to quantize the feature vectors into discrete symbols. This process allows the model to efficiently encode the sequence of symbols into a fixed-length representation for downstream tasks like speech recognition. Subsequently, a contrastive loss function is applied, followed by a diversity loss, enabling the model to learn effective embeddings for speech recognition.

For finetuning, 29 target letters from the Librispeech dataset are utilized. The optimization involves minimizing the CTC loss for ASR tasks, and a modified version of SpecAugment [107] is applied to mitigate overfitting and enhance recognition robustness. Finetuning configurations vary based on the size of the finetuning datasets. The learning rate is adjusted according to the dataset size, and different components of the model are trained sequentially during the finetuning process. The feature encoder is frozen during finetuning, ensuring stability and optimal performance.

It is reasonable to expect that the wav2vec2 models can be adapted for child speech with a small amount of target data. The key advantage of wav2vec2 lies in its ability to learn contextualized speech representations directly from raw audio waveforms. During pretraining, the model captures hierarchical information, including phonetic and linguistic features, without relying on explicit alignment or transcription. This enables wav2vec2 to generalize well to new and unseen data, making it particularly suitable for low-resource scenarios. When adapting wav2vec2 to child speech, the small amount of target data can be efficiently utilized through finetuning. Finetuning involves updating the model's parameters using the limited labelled data from the target domain, in this case, child speech. By initializing the model with the pretrained weights, the finetuning process can quickly adapt wav2vec2 to the specific characteristics of child speech, such as higher pitch, faster speaking rate, and unique pronunciation patterns.

2.5.3 WHISPER

The Whisper model [18] represents a seminal advancement, characterized by its robust handling of a wide range of speech data and its exceptional generalizability. The architecture of the Whisper model is anchored in a transformer-based encoder-decoder

framework. This choice is predicated on the proven scalability and efficacy of such architectures in handling complex sequential data. The audio processing in Whisper involves resampling input to 16,000 Hz and transforming it into an 80-channel log-magnitude Mel spectrogram, calculated over 25 millisecond windows with a 10-millisecond stride. A critical aspect of this processing is the global normalization of input features, aiming to standardize the dataset with near-zero mean values. The architectural specifics (as seen in Figure 7) include an encoder with two convolution layers integrated with the GELU activation function, and the use of sinusoidal position embeddings. The decoder component utilizes learned position embeddings and a unique tokenization strategy that ties input-output token representations, adapted from the GPT-2 model [108], with modifications to accommodate multilingual processing.

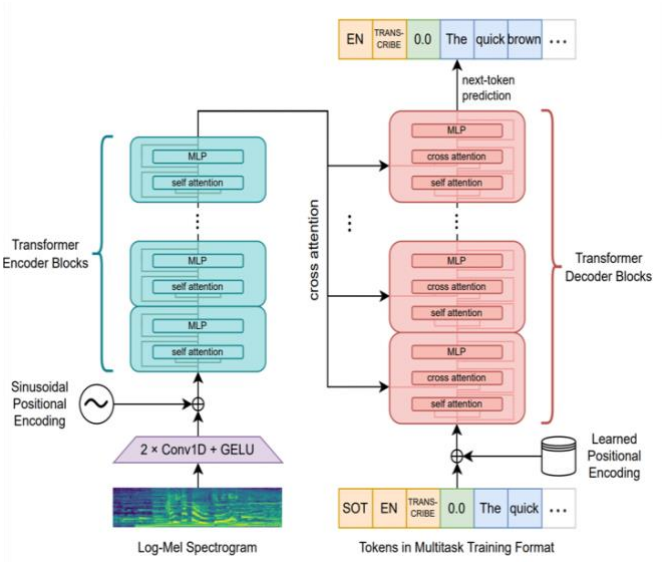


Figure 7: Whisper architecture [18].

The training of the Whisper model is notable for its unprecedented scale and diversity, encompassing 680,000 hours of multilingual and multitask data. This vast corpus of data underpins the model's robust generalization capabilities. The training methodology employed data parallelism across accelerators and leveraged FP16 precision, dynamic loss scaling, and activation checkpointing techniques. The optimization strategy was built around the AdamW optimizer and included a gradient norm clipping [109]. A key to Whisper's training effectiveness was the limited number of updates, strategically chosen to mitigate the risk of overfitting and to emphasize model robustness and generalization, rather than relying on data augmentation or regularization methods. Whisper's multitask format is particularly innovative, extending its functionality beyond transcription to encompass translation, voice activity detection, and language identification. This comprehensive approach simplifies the speech processing pipeline and broadens the model's applicability.

Whisper showed exceptional performance in zero-shot settings, eschewing the need for finetuning on specific datasets, which is a testament to its wide-ranging applicability. The model's performance is characterized by an approach to human-level accuracy and robustness, particularly in out-of-distribution contexts, a notable achievement in the field of ASR. When compared to conventional supervised models, Whisper demonstrates superior robustness, significantly outperforming them across diverse adult speech datasets. For this reason, it was decided to use this model with child speech datasets and to see how

it would behave with unseen and seen child speech datasets. It will be covered in more detail in Chapter 5.

2.5.4 CONFORMER-TRANSDUCER

The Conformer-Transducer [16], a novel approach to ASR, represents a fusion of Transformer and CNNs. This hybrid architecture is designed to harness the respective strengths of both approaches: Transformers' adeptness at capturing global interactions and CNNs' proficiency in extracting local features. Central to the Conformer model (see Figure 8) are the Conformer blocks, which replace traditional Transformer blocks. Each Conformer block consists of four key components: a feed-forward module, a multi-headed self-attention (MHSA) module, a convolution module, and a second feed-forward module. The MHSA incorporates relative sinusoidal positional encoding, enhancing the encoder's robustness to variations in utterance length. The convolution module, starting with a gating mechanism including a pointwise convolution and a gated linear unit (GLU), followed by a depthwise convolution layer, is instrumental in capturing fine-grained local feature patterns. A distinctive feature of the Conformer block is its Macaron-Net-inspired structure. This involves two half-step feed-forward networks (FFNs) sandwiching the self-attention and convolution modules. The use of pre-norm residual units and Swish activation in the feed-forward modules reflects a sophisticated approach to optimizing the network's learning process. These structural innovations improve performance and contribute to the model's parameter efficiency, crucial for large-scale ASR applications.

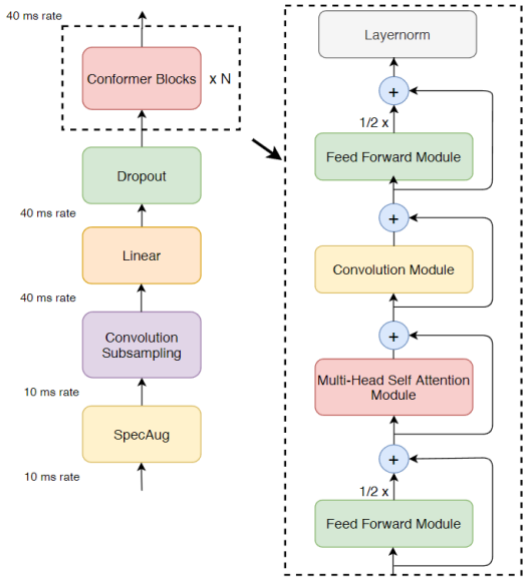


Figure 8: Conformer architecture [16].

Conformer-Transducer offers an improvement in WER for adult speech over the RNN-T and the Transformer architectures. The Conformer-Transducer uses the autoregressive transducer decoder, dropping the original simpler LSTM decoder. It also uses the transducer loss instead of the CTC to reduce incorrect spellings by implicitly telling the model to inherent dependency between predicted output tokens, while CTC assumes that the output tokens are conditionally independent. However, this comes at the cost of larger Graphical Processing Unit (GPU) requirements for training and slower decoding speeds. The Conformer model's integration of convolution within a self-attention framework marks a significant technical innovation in ASR. Its nuanced approach to positional encoding,

optimization of activation functions, and efficient structural design underpin its state-of-the-art performance in ASR tasks.

2.5.5 MODEL PARAMETERS AND SIZES

The Conformer-Transducer, Whisper, and wav2vec2 models each offer a range of versions that cater to different computational needs, trained on datasets varying in size and composition. Table 4 highlights the architectural parameters for wav2vec2, Whisper and Conformer-Transducer models used in this work. The Conformer-Transducer models, available in 'Small' to 'XLarge' sizes, are generally trained on standard ASR datasets like Librispeech[96]. Whisper models show a similar trend, with sizes ranging from 'Tiny' to 'Large'. The 'Tiny' model consists of 39M parameters while the Large model contains 1550M parameters. The wav2vec2 models, with the 'base' and 'large' versions, are distinct for their parameters and training dataset size with the ‘base’ model pretrained with 960 hours and ‘large’ being trained with 60k hours of adult speech.

Table 4: Architecture parameters for Conformer-transducer, Whisper, and wav2vec2 Models

Models	Layers	Width	Heads	Learning Rate	Parameters
Conformer-Transducer Models [16]:					
Small	16	176	4	3.0	14M
Medium	16	256	4	3.0	32M
Large	17	512	8	3.0	120M
XLarge	24	1024	8	3.0	600M
Whisper Models [18]:					
Tiny	4	384	6	1.5×10^{-3}	39M
Base	6	512	8	1×10^{-3}	72M
Small	12	768	12	5×10^{-4}	244M
Medium	24	1024	16	2.5×10^{-4}	769M
Large	32	1280	20	1.75×10^{-4}	1550M
wav2vec2 Models [17]:					
Base	12	768	8	5×10^{-4}	95M
Large	24	1024	16	3×10^{-4}	317M

The effectiveness of each model variant is closely tied to the size and type of dataset it is trained on. The use of these models will become more apparent in Chapter 5. Since we are using these models for improving child speech ASR with limited resources, we maintained most hyperparameters as set by the original authors of each model. This approach not only ensures optimal results but also maintains consistency, which is important for drawing accurate comparisons with other approaches.

Chapter 3

Child Speech and Public Datasets: Challenges and Solution

Before engaging in the technical intricacies of this research, it is important to review the child speech datasets involved. A comprehensive understanding of the characteristics, scope, and constraints of these datasets helps appreciate the challenges in this field of study. The availability of high-quality child speech datasets is a pivotal factor in advancing TTS and ASR technologies for children. Researchers are actively working on creating and curating such datasets to capture the variability within child speech, considering factors like age, language background, and developmental stages. However, these datasets are limited (as will be discussed in section 3.2), especially the ones available for research use.

This chapter provides an in-depth exploration of several key aspects related to child speech. Initially, it addresses the reasons why child speech is considered a low-resource area in research, highlighting the scarcity of data and the unique challenges it presents. A significant portion of the chapter is dedicated to examining the fundamental differences between child and adult speech, which are crucial for understanding the complexities involved in accurately recognizing and processing child speech. The chapter then transitions to a detailed discussion of the specific datasets used in this study, challenges associated with these datasets and methods employed for cleaning and pre-processing the datasets. Finally, the chapter concludes with insights into the development of an application specifically designed for child speech data collection.

3.1 Why Child Speech is a Low-Resource Area of Research?

A low-resource language refers to a language for which there is limited or scarce linguistic data available, especially when compared to more widely spoken or studied languages. This scarcity of resources can include a lack of written or spoken language data, annotated corpora, language models, or language technologies specifically tailored for that language. In child speech research, the focus is often on developing technologies or models that can understand and generate child speech accurately. However, children speak and interact with language differently from adults. They may use simpler grammar, have unique pronunciation patterns, or employ age-specific vocabulary. As a result, developing language technologies for children becomes even more challenging. Therefore, child speech is considered a low-resource area in TTS and ASR research.

The speech and language characteristics of children's voices are substantially different from those of adults [8]. This difference poses a challenge in creating effective models that can accurately recognize and process child speech. Furthermore, there is a notable scarcity of sizable open datasets for children's speech in the research community [110]. This lack of extensive and varied datasets hinders the development and refinement of models

specifically tailored for child speech recognition. Collecting data for child speech research can be a challenging task. Most TTS datasets are created in studios with expensive equipment where an adult will be using a microphone to create clean, noiseless, easy-to-understand, and meaningful audio. This task is not easy to produce and even more difficult to implement with a child.

Child and adult speech also differs significantly in several ways. Children's speech has a higher pitch, slower rate, and less precise articulation due to their developing vocal apparatus. They often use simpler vocabulary and syntax, with exaggerated intonation and less fluent delivery marked by pauses and hesitations. In contrast, adults have a lower pitch, faster speech rate, more precise articulation, and more complex vocabulary and syntax, with controlled prosody and richer voice quality, reflecting fully developed speech organs and greater linguistic experience. These fundamental differences mean that adult speech datasets may not yield comparable results when applied to child speech, making it challenging to achieve accurate models and analysis across different age groups.

Our research also revealed that when plotted in a two-dimensional space, the speech features of adults and children exhibit distinct characteristics. Child and adult speech exhibit distinct characteristics when analysed through speaker embeddings created by a 256-dimensional speaker encoder [131]. These embeddings when plotted in a 2D space showed that child speech embeddings clustered closely together and distinctly from male and female speakers. Cosine similarity measure between adult and child speaker embeddings also showed that adult and child embeddings were quite distant from each other. This will be discussed in more details in Chapters 4 and Chapter 6.

3.1.1 How is Child Speech Different from Adult Speech

Child speech is often described as more noisy or messy compared to adult speech. Children's speech is characterized by rapid developmental changes [5]. As they grow, their vocal tract anatomy and language skills are continuously evolving, leading to significant variability in their speech patterns. This variability makes it challenging to standardize and model child speech for recognition purposes. Children also often exhibit less consistent pronunciation and articulation. Their speech can include mispronunciations, incomplete word formations, and a range of unpredictable variations, unlike the more stable and predictable speech patterns seen in adults [5], [7], [9], [111]. Furthermore, children's speech often includes a higher frequency of non-verbal sounds, such as laughter or crying, and a tendency to mix languages or use idiosyncratic language forms, adding another layer of complexity. In contrast, adult speech tends to be more uniform and stable, with clearer articulation and more predictable speech patterns. Let’s look at the main key differences between adult and child speech can be broken down into several areas as follows:

Physiological Differences [112]: The anatomical structure of children's vocal tracts is significantly smaller than that of adults, which leads to distinct differences in prosody features [8], [113]. This size disparity results in children's voices having characteristics and features that are markedly different from those of adults. For instance, the fundamental frequency or pitch of children's speech is considerably higher [83], [114]. While adult speech typically ranges from 70 to 250 Hz in pitch, children's speech pitch spans between 200 to 500 Hz [113]. This higher pitch is attributed not only to their smaller vocal cords but also to their overall smaller body size, which influences the resonance and timbre of their voice. Additionally, there are noticeable differences in the speaking rate [114], [115]. In children speech, the average duration of each phoneme is longer, leading to a slower

speaking rate compared to adults. This slower rate of speech in children is not solely due to their smaller vocal cords but is also influenced by various factors such as their developmental stage, cognitive processing abilities, and linguistic skill level. As children grow and develop, their speech rate, pitch, and other prosodic features evolve, adding another layer of complexity to the task of accurately recognizing and analysing a child's speech. This evolution in speech characteristics presents unique challenges in speech recognition and processing, requiring specialized models that can adapt to the changing features of a child's voice.

Linguistic and Developmental Differences: The speech of children is marked by greater variability than that of adults. A multitude of factors, including the child's age, stage of development, and individual speech traits, lead to a broad spectrum of speech patterns [116], [117]. This extensive variability presents considerable challenges in both modelling and comprehending child speech. As children are in the developmental stages of acquiring and honing their speech abilities, they frequently exhibit unclear articulation and significant variability in their pronunciation. The vocabulary of a child is generally more constrained compared to adults, and their use of sentence structures tends to be simpler. In the process of mastering language, children are still grasping the intricacies of grammar and syntax, aspects that are discernible in their speech patterns. This ongoing development in language skills means that their speech can change rapidly over a short period, reflecting their learning curve. This aspect of speech evolution in children adds a dynamic dimension to their speech, making it a moving target for speech recognition systems [3], [4]. Understanding and accommodating these developmental differences are crucial in creating effective models for child speech recognition, as they need to be flexible enough to adapt to the evolving nature of a child's linguistic capabilities.

Behavioural and Environmental Differences: Children's speech can be less predictable and more spontaneous as compared to adults. Children frequently alter their speaking pace or loudness unexpectedly, and their speech is often imbued with a wide range of emotions and expressions [14], [118]. This spontaneity can pose a challenge for consistent speech recognition. Additionally, recordings of children's speech are typically characterized by a higher level of background noise. Children often speak in less structured settings, such as play areas or classrooms, where a variety of extraneous sounds are present. These environments can introduce a multitude of unpredictable acoustic elements into the recordings, complicating the task of isolating and analysing the child's speech. This combination of behavioural unpredictability and environmental noise adds layers of complexity to the process of capturing and processing child speech data effectively.

Increased Acoustic Variability and Rapid Evolution of Speech: The acoustic and linguistic properties of children's speech are significantly more varied and unpredictable compared to adults. This is due to various factors such as age, developmental stage, and individual differences in speech patterns, including pronunciation, speed, and prosody [8], [118], [119]. These elements contribute to a heightened level of complexity in accurately recognizing and processing child speech. Furthermore, as children grow, their speech undergoes rapid changes. The dynamic nature of their speech development means that datasets capturing their speech can quickly become outdated. Therefore, there's a continual need to update these datasets to keep pace with the evolving speech characteristics of children, further increasing the complexity and resource demands in child speech research.

3.1.2 What Are the Technical Challenges Associated With Child Speech Research?

The field of child speech research is riddled with unique technical challenges, significantly different from those encountered in adult speech processing. These challenges stem from various factors, including the nature of child speech, data collection constraints, and the evolving ethical landscape in research. Each of these factors contributes to the complexities of developing accurate and reliable speech recognition and synthesis systems for children. These factors include:

Limited Data Availability [110], [120]: Unlike adult speech data, which can be readily sourced from various mediums like studios, YouTube, audiobooks, etc., collecting child speech data presents unique challenges. The primary issue stems from the environments in which child data can be ethically and legally collected. Unlike adults, recording child speech often requires specific conditions, such as educational settings or controlled environments, to ensure safety and comfort. Furthermore, regulations such as the GDPR [41], [45], [121] impose strict guidelines on how personal data, especially of minors, can be collected and used, adding layers of complexity to the data collection process. These factors collectively lead to a significant limitation in the availability of diverse and extensive child speech datasets.

Limited Expertise: There is a limited pool of researchers and developers with the specialized knowledge required to tackle the unique challenges of child speech recognition and synthesis. This lack of expertise further contributes to the area being under-resourced.

High Annotation Costs: The annotation of child speech data is a complex and nuanced process, requiring specialized expertise in child language development. Children's speech is characterized by pronunciation errors, evolving grammar, and simplified language structures, making their speech fundamentally different from adults. This complexity necessitates a detailed and careful approach to annotation, often involving multiple revisions to capture the nuances of a child's linguistic development accurately. This specialized process is both time-consuming and costly, as it requires linguists with specific expertise in child language. Consequently, the scarcity of accurately annotated child speech data poses a significant challenge in developing effective speech recognition technologies for children, underlining the need for more focused resources in this area.

Lack of Standard Benchmarks: Unlike adult speech, where there are well-established benchmarks and datasets, child speech lacks such standardization. This absence of benchmarks makes it difficult to measure progress and compare different systems for child speech. This absence hinders the ability to track developmental progress, establish standardized practices, and conduct comparisons across various studies or system implementations.

In light of these challenges, our research endeavours to navigate and address these complexities. We utilize publicly available datasets and have developed a comprehensive cleaning and pre-processing methodology. This approach is designed to enhance the usability and standardization of these datasets, making them more suitable for research in the context of the unique characteristics of child speech.

3.2 Datasets Used in This Study

The nature of this study, considering the challenge of limited children's speech datasets and the multi-step training process involved, calls for the use of multiple large datasets, including adult speech datasets. The specific applications of these datasets will become increasingly apparent as per their usage in subsequent chapters. Each dataset has been selected for its unique suitability, each contributing distinctively to the different experimental setups and objectives. This structured approach allows for a nuanced exploration of the multifaceted aspects of child speech research, highlighting the versatility and importance of each dataset in the broader context of our study.

3.2.1 Adult Speech Datasets Used in This Study

In the context of our research focusing on ASR and TTS technologies for child speech, we employed a comprehensive range of adult speech datasets. These include Librispeech[96], LibriTTS[122], VCTK[123], VoxCeleb1[124], LJ Speech[78], and Librilight [125]. Utilizing these diverse datasets allows us to access a wide variety of speech patterns, accents, and linguistic nuances (will be discussed in more detail in Chapters 4 and 5). This variety is crucial for training and refining our ASR and TTS models, ensuring they are robust and versatile enough to handle the complexities of child speech. The demographics of these datasets are made available in Table 5.

Table 5: Adult Speech Datasets Used in This Study

Dataset	Speakers	Hours	Comments
Librispeech [96]	2400	960	English speech dataset derived from audiobooks. Popularly used in ASR research.
LibriTTS [122]	2400	585	TTS dataset derived from Librispeech corpus.
VCTK [123]	110	44	Recordings from various English accents and are highly used in multi-speaker TTS research.
VoxCeleb1[124]	1251	352	Celebrity voices extracted from YouTube containing 153,516 utterances from 1,251 speakers.
LJ Speech [78]	1	24	Popularly used in TTS research
Librilight [125]		60,000	Self-supervised dataset of adult speech containing 60k hours of audio files without transcription. Popularly used in SSL ASR training.

Adult speech datasets, including LJ Speech, Librispeech, LibriTTS, VCTK, and VoxCeleb1, are essential in ASR and TTS research for child speech. Utilizing adult speech datasets alongside child speech datasets can significantly improve research outcomes for child speech in ASR and TTS. Adult datasets provide a robust and diverse foundation of speech patterns, which can be leveraged to train and develop initial models. These models, once trained on adult speech, can undergo transfer learning processes, adapting their learned features to cater to the nuances of child speech. This approach not only broadens the model's understanding of speech variability but also compensates for the gaps and limitations inherent in child speech datasets. Furthermore, in ASR systems, the inclusion of adult datasets adds a layer of phonetic and prosodic complexity, enriching the model's capability to process the broader spectrum of speech characteristics found in child speech.

Additionally, these adult datasets are instrumental in cross-age testing, evaluating the performance and adaptability of models across different age groups. This testing is crucial for understanding the effectiveness of child speech models on adult speech and vice versa, ensuring the development of versatile and adaptable speech technologies.

3.2.2 CHILD SPEECH DATASETS USED IN THIS STUDY

We also utilized various child speech datasets such as the My Science Tutor (MyST) corpus [126], PFSTAR [127], CMU_Kids [128], and speechocean762 [129]. These datasets are pivotal in providing a rich source of data that encapsulates the unique linguistic characteristics and speech patterns of children. The MyST dataset, for instance, offers a diverse range of child speech recordings in an educational context, making it invaluable for understanding and modelling how children interact in learning environments. PFSTAR and CMU Kids contribute significantly with their varied samples of child speech, encompassing different age groups, accents, and dialects. The speechocean762 further enriches our research resources with its collection of child speech data in non-native Chinese accents. The utilization of these child-specific speech datasets ensures that our research is grounded in real-world speech characteristics of children, allowing for the development of more accurate and effective speech technology solutions tailored for younger users. More details about these datasets can be seen in Table 6.

Table 6: Child Speech Datasets Used in This Study

Dataset	Age Range	Speakers	Hours	Comments
My Science Tutor (MyST) corpus [126]	Grade 3-Grade 5	1371	393	Advantages: ▪ The largest corpus of child speech dataset available open source for research use. ▪ 49% of the data is transcribed. Disadvantages: Contains a lot of noisy data
PFSTAR [127], ▪ British English ▪ German ▪ Italian ▪ Swedish	6-11 10-15 9-11 4-8	159 57 78 40	14.1 3.4 3.5 1.2	Contains child speech including read and spontaneous native and non-native dialects in British, German, Italian and Swedish accented English. Clean speech in comparison to other child speech datasets.
CMU Kids [128]	6-11	76	9	Very noisy dataset, and difficult to understand.
Speechocean762 [129]	6-43	250	6	Consists of 5000 English utterances from 250 non-native Chinese (Mandarin) speakers, where half of the speakers are children

It's evident from Table 5 that the quantity of adult speech datasets significantly exceeds that of child speech datasets as listed in Table 6. Let's delve deeper into the difficulties linked with these child speech datasets and explore potential solutions for overcoming these challenges.

3.2.3 PROBLEMS ASSOCIATED WITH THE CHILD SPEECH DATASETS USED IN THIS RESEARCH

We conducted a comprehensive study on child speech datasets. With the aim of maintaining a standardized methodology, we analysed both the audio recordings and their corresponding transcriptions. The goal was to determine the most effective strategies for cleaning and refining the datasets to ensure the highest level of clarity and usability. This involved a detailed examination of the transcriptions and various intricacies present within these datasets, as well as identifying the challenges they pose. Due to the absence of corresponding transcripts for many audio files within these datasets, the decision was made to exclude such audio files from the dataset. For the audio files with transcripts in hand, we observed a lot of transcripts without phonetic meaning and were missing quite some pronunciations. The transcripts of some example audio files from the MyST dataset [126] are listed below in Table 7 to illustrate some of the problems. It was also important to establish a standardized naming and saving convention for the audio files and their transcriptions, tailored to the specific ASR/TTS training methodologies being employed. Therefore, It was decided to clean these datasets to make them usable for TTS and ASR training as discussed in section 3.2.4.

Table 7: Problems Seen in Transcripts of the MyST Child Speech Dataset [126]

Problems identified in Transcripts	Example
Audio files containing noise in their utterances without any phonetic meaning	▪ myst_004029_2013-12-17_09-34-22_EE_2.2_009.wav| <noise> ▪ myst_002268_2015-04-21_08-14-11_LS_3.1_013.wav| it's glowing <breath>
Audio files that are not coherent or indiscernible.	▪ myst_002267_2015-04-30_13-20-05_LS_3.3_014.wav| in oxygen right <indiscernible> ▪ myst_004029_2013-12-03_09-52-03_EE_1.4_016.wav| ‘‘can hear sound because of that <indiscernible>’’
Audio files are too small in length	▪ myst_002013_2014-03-11_11-14-16_LS_2.1_025.wav | ‘‘energy <noise>’’
Audio files are too long	▪ myst_002013_2014-03-11_11-14-16_LS_2.1_014.wav | ‘‘it’s trying to show us that all the things that it needs all the things that the plants needs to grow it needs soil on the bottom it needs at least a ground a top the a a top to lay on for the plant to grow so you can see it that’s only with flowers and plants it’s not with vegetables and it needs and it needs the energy from the sunlight to grow and it needs water because somebody’s watering the plant.’’
Transcription containing text with no phonetic information.	▪ myst_002033_2014-03-10_13-45-32_LS_2.1_020.wav | ‘‘(())(())(())’’ ▪ myst_002030_2014-04-30_10-32-58_LS_4.2_011.wav| the (()) (()) might help it push
Repetition of words/stammering noticed in children’s voices.	▪ myst_990027_2008-21-04_00-00-00_MS_1.1_050.wav | ‘‘um we measured how big a millimeter meter is a meter and a kolome- a ∗ kilometer ∗’’

	▪ myst_990027_2008-21-04_00-00-00_MS_1.1_046.wav|it was uh a cuh- uh cold between warm day it's col- cold mostly cold

In the course of our research, numerous issues were identified in the audio files of the child speech datasets, beyond just the accuracy of the transcripts. A significant portion of these audio recordings exhibited various forms of noise, such as distortions from children speaking too close or too far from the microphone, ambient background noise, stammering and missing words, as well as instances of extremely low or high volume. Some files even contained sounds with no phonetic significance, amounting to pure noise. To address these challenges, these problematic files were meticulously removed from the datasets. Additionally, several datasets were segmented into smaller sections to enhance their utility for training purposes. It is also important to note that for uniformity and ease of use in TTS/ASR training, all the training audio files were converted into the .wav format, while the transcription files were standardized in the .txt format. Detailed methodologies with regards to the cleaning and pre-processing specifics for some of these datasets are elaborated in the ASR and TTS publications, which are included in Appendix B, C and E of this thesis.

3.2.4 Cleaning and Preprocessing of Child Speech Datasets

To preprocess and clean child speech datasets for TTS and ASR research, a comprehensive approach was adopted, drawing from methodologies used in the development of the LibriTTS dataset [122]. Audio files are first standardized: for ASR, they are converted to 16-bit depth at a 16 kHz sampling rate, and for TTS, a 24 kHz sampling rate is used, using tools like ffmpeg[1] and sox[2]. Afterwards, text data normalization is performed, where abbreviations and punctuation are systematically replaced. This normalization extended to whitespace and character case uniformity. Non-linguistic annotations in the datasets, such as various symbols and noise markers, were meticulously removed, retaining only alphanumeric characters in the transcripts. These non-linguistic content (in child speech datasets) included annotations such as ‘‘<unk>, sil, hmm, <breath>, <noise>, <indiscernible>, [ze-], [cham-], [∗∗∗ision], etc.’’ which were also removed through basic text processing tools.

The dataset was then segmented into smaller, more manageable chunks. This was achieved through the use of the Montreal Forced Aligner[3], which provided forced alignment of the speech data. This alignment process generated precise timestamps, correlating transcripts with their corresponding audio files at the word and phone levels. These timestamps were then utilized to segment larger audio files into smaller sections, typically ranging from 5 to 25 seconds. Longer audio files that proved challenging for segmentation or were not meaningful, as well as shorter files typically filled with noise, were excluded from the dataset.

This thorough cleaning and preprocessing protocol ensured the creation of a dataset that was not only cleaner but also more conducive to research, specifically tailored to the unique requirements of child speech data in TTS and ASR applications. The focus on dataset quality and usability was crucial in addressing the inherent challenges posed by child speech, paving the way for more effective and efficient speech recognition and synthesis

[1] https://ffmpeg.org/
[2] https://github.com/chirlu/sox
[3] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

research. This allowed us to create a cleaner and more research-friendly version of each of the child speech datasets used in this work. The detailed demographic on cleaner subsets of these child speech datasets is provided in Table 8.

Table 8: Child Speech Datasets Demographics (Post-Cleaning)

Dataset (post-cleaning)	Hours	Comments
MyST Original MyST Subset: ▪ TinyMyST ▪ MyST_2 o MyST_train/MyST_55h o MyST_test	393 19.22 55 10	▪ TinyMyST is a subset of MyST containing clean, transcribed speech used in our TTS experiments. ▪ MyST_2 is a subset of MyST which contains 65 hours of clean child speech. This dataset contains parts of TinyMyST as well as an additional dataset from the untranscribed MyST subset which was transcribed using methodologies discussed in section 3.2.4. ▪ MyST_2 is divided into MyST_train (55hrs) and MyST_test (10hrs). MyST_train is used in most of our experiments for training TTS and ASR models. MyST_train is also referred to as MyST_55h in ASR experiments.
PFSTAR ▪ British English o PFS_train/PFS_10h o PFS_test ▪ German ▪ Italian ▪ Swedish	 12 10 2 3.4 3.5 1.3	▪ Different PFSTAR subsets were segmented and combined as per their usage in ASR experiments. ▪ British English is referred to as PFS in our experiments since it was the most used dataset from the PFSTAR dataset. It was divided into PFS_train and PFS_test for most experiments. ▪ German, Italian and Swedish subsets were only used for Non-Native child speech ASR experiments and are covered in more detail in Appendix E.
CMU Kids ▪ CMU_Kids ▪ Non-Native Experiments o CMU_train o CMU_test	9 9 7 2	▪ Transcripts were cleaned to remove non-meaningful and non-phonetic information. ▪ CMU Kids is mostly used as an unseen inference dataset for ASR experiments (named CMU_Kids for those inferences). ▪ It was divided into CMU_train (7hrs) and CMU_test (2hrs) for Non-Native child speech ASR experiments (see Appendix E).
Speechocean762	2.4	▪ We selected speakers whose ‘Age’ was less than ‘18’, amounting to 2.4 hours of child speech.

The dataset created using the described cleaning methodology illustrates a well-structured and specialized approach to child speech dataset preparation for TTS and ASR research. This approach, with attention to detail in segmentation and transcription, significantly enhances the dataset’s utility for research, paving the way for more accurate and efficient child speech technologies. These datasets, now in a cleaner and more structured form, are tailored to effectively support the various TTS and ASR experiments detailed in our study.

3.2.5 TRAINING DATA REQUIREMENT FOR ASR AND TTS SYSTEMS

The development of speech technologies, ASR and TTS, relies heavily on extensive and well-structured training data. Understanding the distinct data requirements for these systems is crucial for optimizing their performance and achieving high levels of accuracy and naturalness. This discussion highlights the key differences in training data requirements for ASR and TTS systems, particularly in terms of the number of speakers, the amount of data per speaker, and the recording conditions, with a focus on child speech datasets and adult datasets used in the field.

ASR systems require a large number of speakers to ensure diversity and robustness of the model. Typically, these systems utilize datasets with thousands of speakers to capture the variability in pronunciation, accents, and speaking styles. For instance, the Librispeech [96] dataset contains approximately 2,400 speakers and provides 960 hours of speech, averaging roughly 24 minutes of data per speaker. Conversely, TTS systems can achieve effectiveness with fewer speakers, as demonstrated by datasets like LJ Speech [78] (24 hours) and LibriTTS [122] (585 hours), emphasizing quality and consistency over sheer volume.

When considering data per speaker, ASR systems prioritize short utterances from numerous speakers to compile a substantial dataset. ASR systems generally require more data per speaker to build robust acoustic models that can handle the diversity of speech patterns. In child-specific datasets like MyST_Complete [126], the average data per speaker is 15 minutes, with individual contributions ranging from as low as 1.72 minutes to as high as 110 minutes per speaker. The inclusion of cleaned subsets of child speech like TinyMyST and MyST_2, with the abundant adult speech datasets enhances ASR experiments. Conversely, TTS systems may require a smaller number of speakers, as the focus is on generating high-quality synthetic speech from a single or a few target voices. The effectiveness of TinyMyST and MyST_2 datasets is evident in TTS experiments (to be detailed in Chapter 4), showcasing the ability to produce good-quality child speech despite their modest size. Each TTS dataset needs to be comprehensive, covering various phonetic contexts, regardless of data per speaker. For example, the TinyMyST dataset spans speech durations from 10.2 seconds to 8 minutes per speaker, contributing to a comprehensive 20-hour dataset. Combined with finetuning techniques, these datasets played a pivotal role in achieving optimal TTS results.

Recording conditions also play a crucial role in shaping dataset characteristics. ASR systems must accommodate diverse real-world conditions, including background noise and varying microphone qualities. In contrast, TTS systems require clean, noise-free recordings, often produced in controlled studio environments. The Librilight dataset [125] exemplifies ASR-focused recordings, capturing diverse conditions to enhance model robustness, while the VCTK corpus [123] provides TTS-specific recordings under controlled conditions to ensure natural and intelligible synthetic speech.

In summary, ASR systems benefit from a large number of speakers and can tolerate diverse recording conditions, whereas TTS systems require fewer speakers but need extensive, high-quality data per speaker recorded in controlled environments. These differences reflect the distinct objectives of each technology as ASR aims for broad recognition accuracy, while TTS focuses on generating high-quality, natural-sounding speech.

3.3 Building an Application for Child Speech Data Collection

In the context of the DAVID initiative, our efforts extended to the development of an interactive application designed for children's engagement. The application presents sentences for the children to articulate, thereby serving a dual purpose: to facilitate an interactive learning experience and to gather child speech data in collaboration with Xperi data collection. This endeavour aligns with our project goals for data collection with Xperi for the development of the DAVID smart toy. The app was developed with the invaluable expertise of Joe Desbonnet, who has over 30 years of experience in software engineering and entrepreneurship. Joe's is a postdoctoral researcher in our research group and his contributions were significant in crafting an engaging and effective application.

The development of the app for child speech data collection underwent a comprehensive process, commencing with a meticulous design phase. During this phase, the development was concentrated on crafting the user interface (UI) of the app, carefully selecting a colour scheme and graphics that would be visually appealing and engaging for children. The UI was engineered to be intuitive and child-friendly, ensuring that the children could navigate it with ease and interest. Attention was also given to the cognitive load of the app, making sure that it was age-appropriate and did not overwhelm the young users. Functionalities such as interactive prompts and feedback mechanisms were integrated to foster an immersive learning environment.

Following the design phase, the project moved into the Proof of Concept (PoC) testing. This critical phase involved real-world testing of the app with children to validate the design choices and functionalities. We initially reached out to colleagues within our network who had children, inviting them to participate in the testing phase. This approach allowed for a controlled yet authentic testing environment where children could interact with the app in a natural setting. Feedback from these initial users was invaluable, highlighting areas for enhancement such as the responsiveness of the app, clarity of instructions, and the overall user experience. Based on this feedback, iterative improvements were made to refine the app.

With the successful completion of the PoC testing, the app was then ready to be deployed for actual data collection. This final stage marked the transition from a controlled testing environment to a broader and more diverse field of use. The app, now finetuned and validated, was utilized to present sentences to children, who would then articulate them, allowing for the capture of a rich array of child speech data. This will be discussed in more detail in section 3.4.

3.3.1 TECHNOLOGIES INVOLVED

The app's technical infrastructure was built on Tomcat[1] web apps, providing a robust Java HTTP web server environment. MariaDB[2] was employed for MySQL support, ensuring a reliable database system for storing and managing the collected data. Additionally, the application was containerized using Docker[3], which streamlined deployment and scalability. The app's current iteration is available privately on GitHub. Furthermore, the integration of the Mozilla Web Speech API[4] offered a seamless speech recognition

[1] https://tomcat.apache.org/
[2] https://mariadb.org/
[3] https://www.docker.com/
[4] https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API

interface, accessible directly from web browsers, which was pivotal in facilitating the speech data collection from children using the web app.

3.3.2 WORKING OF THE APPLICATION

Figure 9 presents the workflow of a web-based application for collecting child speech data. In this process, a child interacts with the application by reading aloud an English sentence displayed on the web browser. This spoken input is recorded and subsequently analysed by an integrated cloud-based ASR system, which transcribes the audio and evaluates it against the original text. The transcription's accuracy is scored using the WER metric—a standard assessment tool in speech recognition. If the transcription's accuracy surpasses a set threshold, indicating a high-quality speech sample, it is stored anonymously in a database. Conversely, if the ASR score falls short of the threshold, indicating potential errors or unclear speech, the system still retains the sample but prompts the child to repeat the sentence. This ensures a comprehensive collection of data, capturing a wide range of speech clarity.

The Harvard sentences[1] are employed as input for this application, enabling children to speak these sentences and record them, with the recordings being securely stored in a protected server. Harvard sentences are carefully crafted to be phonetically balanced. This means they contain a wide range of phonemes (basic units of sound in a language) in proportions that are typical in everyday speech. For children, these sentences are particularly useful as they cover a broad spectrum of sounds found in the English language, which is beneficial for speech recognition and synthesis systems. These sentences are designed to be easily spoken, making them suitable for children who are still developing their speech and language skills.

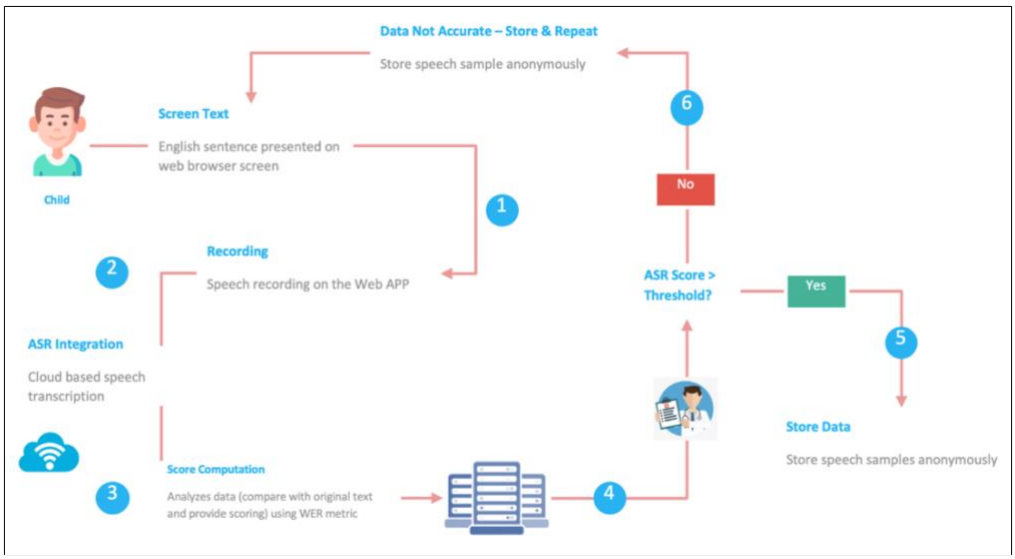


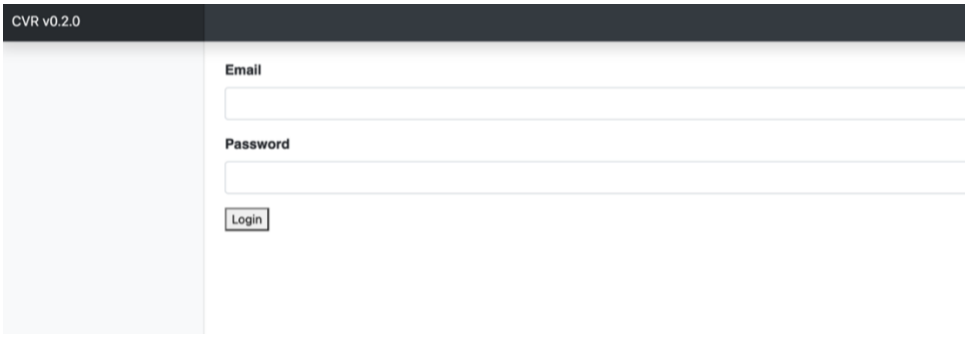
Figure 9: Flow diagram showing the working of the child speech data collection application.

3.3.3 APPLICATION INTERFACE

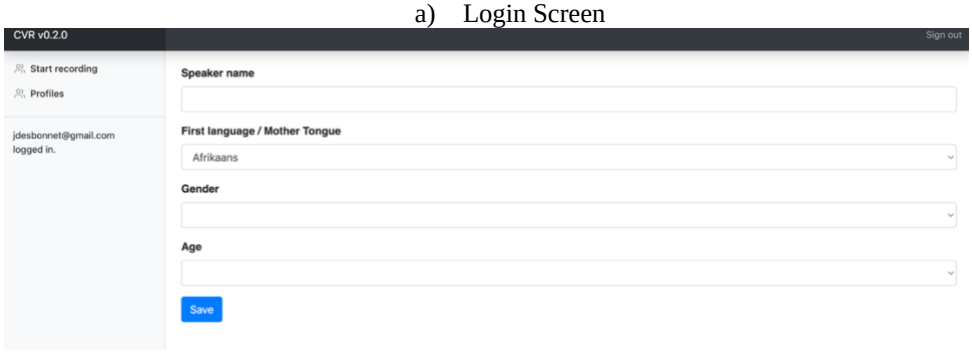
The application starts with a user login interface, which is designed to be child-friendly, ensuring ease of access for young users. This step also involves parental permission, considering the app is used by children. Harvard sentences are used to provide input to the app. The core functionality of the app involves recording children's voices as they speak

[1] https://harvardsentences.com/

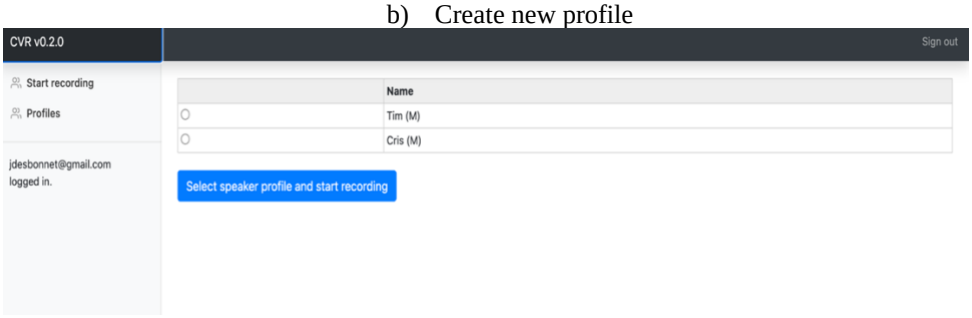
the Harvard sentences. The interface for this is made visually engaging and intuitive, encouraging children to speak naturally and comfortably. After recording, the application stores the voice data for processing. This involves uploading the data to a server for further analysis. Figure 10 illustrates these different elements of the application, providing a visual guide to its usage.



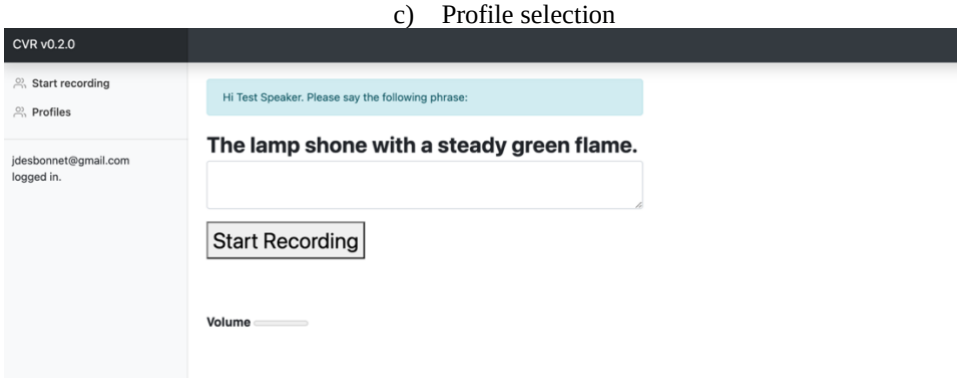
a) Login Screen



b) Create new profile



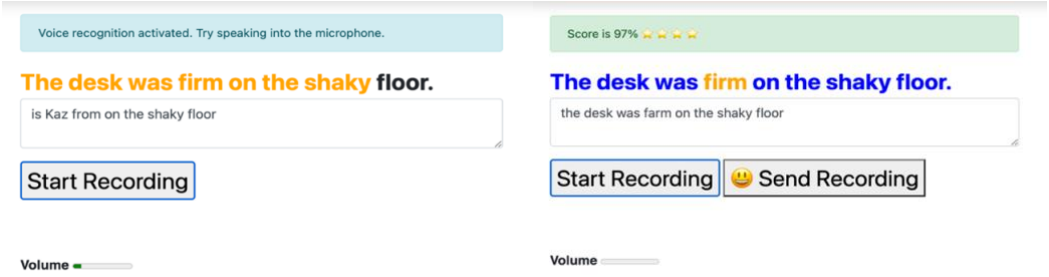
c) Profile selection



d) Recording Screen

Figure 10: Child voice recording application interface.

The speech recording feature in the application employs a colour-coded feedback system to indicate pronunciation accuracy. When a child's speech is detected as incorrect, the text colour changes to yellow, while correct pronunciation results in a blue text colour. For visual representation, refer to Figure 11.



a) Output: Incorrect pronunciation b) Output: Correct pronunciation

Figure 11: App output for incorrect and correct pronunciation.

The data recorded by the app was stored in a JSON format containing important metadata such as ID, age, transcription, speaker, language etc. associated with the child speaker. An illustration of this format can be seen in Figure 12. It also contains a base64 audio representation which can be easily exported to .wav audio codec. In the image below, the field highlighted in white represents the base64 Audio encoding. The audio files are also stored in .mp3 and .wav format along with the .json metadata files.


```
{"id":149,
"timestamp":"Apr 19, 2023, 4:11:16 PM",
"name":"3840","browserAgent":"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36",
"transcription":"the red paper brighten the dim stage",
"originalPhrase":"The red paper brightened the dim stage.",
"score":0.947368421052631 5,
"audioData":"data:audio/mp3;base64,GkXfo59ChoEBQveBAULygQRC84EIQoKEd2VibUKHgQRChYECGFOAZwH/////// FUmpZpkq17GDD0JATYCGQ2hyb21lV0GGQ2hyb21lFlSua7
"userId":1,
"speakerId":20,
"gender":"M",
"age":7,
"firstLanguage":"EN"}
```


Figure 12: Example of a JSON file storing the audio metadata collected using the application.

3.4 Exploratory Child Speech Data Collection Activities

The development and deployment of a specialized app for child speech data collection, as part of the DAVID project, was an initiative undertaken by the University of Galway[1] to collect child speech data with Xperi[2] in Ireland and later extended to BITS Pilani[3] in India. This app serves as a cornerstone in addressing a critical need in the research field of child speech: the creation of clean, usable, and diverse child speech datasets. At Xperi, the app's deployment was driven by the necessity to develop AI technologies that could interact effectively with children. Recognizing that children's speech patterns are distinctively different from adults', the app was designed to capture a wide range of speech data in various interactive scenarios. This data is invaluable, not just for enhancing AI's ability to understand and respond to children's speech, but also for contributing to the broader understanding of child language development and linguistics. The application's use at BITS Pilani further expanded its impact. Here, the focus was on capturing the unique characteristics of Indian-accented English among children, reflecting the region's rich linguistic diversity.

3.4.1 Data Collection at Xperi, Galway

As a part of the DAVID project, Xperi was involved in the extensive collection of audio and visual data from children. The participants ranged in age from 3 to 12 years. These

[1] https://www.universityofgalway.ie/c3i/projects/david/
[2] https://xperi.com/
[3] https://www.bits-pilani.ac.in/

young individuals were also encouraged to interact with developmental toys for feedback collection. Different cameras and recording devices were employed to capture the speech and visual features of children in diverse settings. This encompassed the 3D data collection, capturing the speech of children interacting with childcare professionals, toys, and applications, among other scenarios.

Figure 13: Xperi fullbody 3D scanner.

The visual data acquisition process for this features a photogrammetric 3D scanner. This device is capable of capturing a series of high-resolution images of subjects or objects placed within its confines. Subsequently, these images are processed to create a detailed 3D model of the subject. Xperi has developed software which not only processes the images into 3D models but also employs algorithms to animate these models. Figure 13 illustrates the 3D scanner. Children were instructed to step into the 3D scanning environment (see Figure 14) and engage with an application designed to recite and enunciate Harvard sentences. The purpose of this activity was for the children to listen and then verbally repeat these sentences aloud. Following this interaction, a comprehensive 3D scan of each child was conducted using the 3D scanner.

Figure 14: Example images of children in the 3D scanner room (images from Xperi data acquisition).

In addition to the photographic data, there's also an initiative to collect speech data. The child speech data collection was a joint effort by Xperi and the University of Galway, designed to efficiently capture diverse speech patterns in children. The project was crafted collaboratively by researchers from both entities, combining Xperi's technical expertise in audio processing with the University of Galway's academic knowledge in linguistics and child development. Speech data from the children was meticulously gathered in a variety of recording settings, employing microphones strategically positioned around the room (see Figure 15). These environments were designed to capture the nuances of the children's

speech as they interacted with a range of elements such as smart toys, childcare workers, games, and various applications. This setup ensured a comprehensive audio capture, reflecting the natural variations in the children's speech across different contexts and activities. The strategic placement of microphones throughout the room played a crucial role in obtaining a clear and accurate record of the children's verbal expressions and interactions in these diverse scenarios.

Figure 15: Xperi data collection playroom environment.

One specific application, integral to Xperi's data collection protocol, facilitated interactions with children to record their speech in a reading environment. The application developed in the previous section (section 3.3) was employed for this task. This read-speech scenario was particularly designed to capture the children's spoken responses to prompts or passages displayed on the application. The children were asked to read aloud from the app, which not only allowed for the collection of speech data in a controlled setting but also provided valuable insights into their reading and speech patterns. This was captured by many different microphones in different environmental settings. This scenario was instrumental in understanding how children of different ages and developmental stages interact with technology and respond to guided speech tasks. The presence of a childcare worker was crucial in this setting as well, ensuring that the children remained focused and engaged, thereby facilitating a more natural and effective data-collection process.

3.4.2 DATA COLLECTION AT BITS PILANI - INDIA

The application (from section 3.3) is currently being utilized in an ongoing research collaboration project at BITS Pilani in India. Here, a dedicated team of researchers employed the app as a vital tool in their project aimed at gathering a dataset of Indian-accented English spoken by children. This endeavour was part of a broader research initiative to analyse and understand the nuances of English speech patterns among Indian children, which often carry distinctive regional accents. Their objective is to compile a rich and diverse dataset that accurately reflects the variations and complexities of English as influenced by India's varied linguistic landscape.

Chapter 4

Contribution to Improving TTS Technologies for Child Speech

This was our first extensive experimental work carried out as a primary task for this thesis using neural TTS algorithms. One of the key responsibilities undertaken by the University of Galway in the DAVID project [24] was to provide support to Xperi in the development of TTS synthesis using Edge-AI technology. Initially, this entailed collaborating on state-of-the-art TTS models to assess their performance. However, as the project progressed, a new objective emerged: to expand the child-speech training dataset for TTS systems. The rationale was that Edge models require optimization, and having access to more extensive training data can significantly enhance their performance. Consequently, this shift in focus led to a dedicated research initiative aimed at enhancing TTS technology for child speech. Therefore, the study focuses on building a viable pipeline for the synthesis of children's voices with low data requirements, which could also facilitate the creation of large synthetic datasets to support other child speech research areas like ASR and speaker recognition [44], [130].

4.1 Tacotron 2-based Transfer Learning Methodology for Child Speech Synthesis

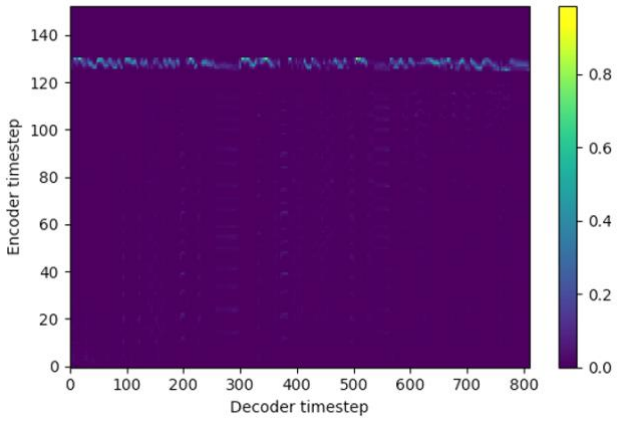
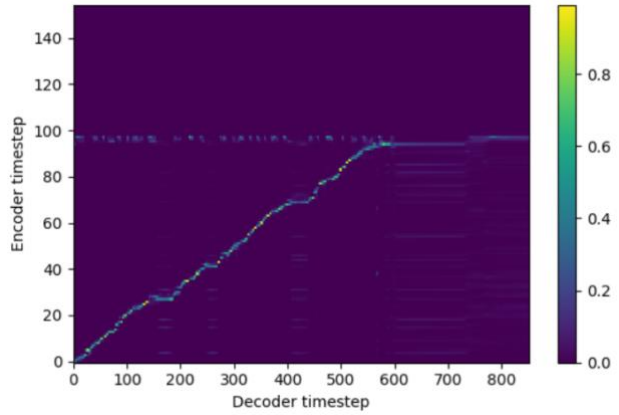
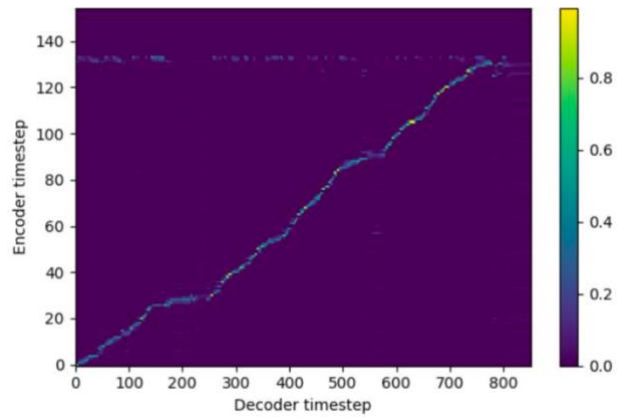
Our contributions in this work were tailored towards the development and validation of a training pipeline for finetuning neural TTS models using child speech datasets, which involved, cleaning a publicly available child speech dataset to provide a usable subset of approximately 19 hours, implementing a multi-speaker TTS retuning workflow for transfer learning and performing both subjective and objective evaluations to assess the TTS model's performance, showing strong correlation and similarity between real and synthetic child voices. This was the first published research work highlighting the contributions to TTS for child speech.

4.1.1 Single Speaker TTS Training

While Tacotron 2 [28] is recognized as a leading system for TTS, its standard configuration is optimized for datasets featuring a single speaker, such as the LJSpeech dataset [78], thereby limiting its output to the vocal traits of the single speaker represented in the training data. When Tacotron 2 was released, it served as a SOTA and still is being used to draw comparisons with newer research. In our preliminary experiments, various SOTA TTS models (as mentioned in Table 3) were subjected to training using a transcribed segment of the MyST dataset called TinyMyST, but the results obtained didn’t have any meaningful audio output.

In TTS, alignment refers to the mapping between the input phonetic or linguistic representation and the acoustic output. In sequence-to-sequence models with attention, such as Tacotron, the alignment is learned by the model so that each timestep in the output sequence (e.g., a frame of audio) corresponds to the appropriate timestep in the input sequence (e.g., a phoneme or word). An alignment plot can be used to visualize this alignment. Ideally, you would expect to see a clear diagonal line across the plot, indicating that each part of the input corresponds to a successive part of the output. Deviations from the diagonal can indicate issues such as mispronunciations or unnatural prosody.

Table 9: Alignment Plots for Different Single-Speaker Tacotron 2 Training Experiments

Training Types	Alignment plots	Comments
I. Tacotron 2 trained with the Original MyST dataset for up to 200k steps		Missing Alignment indicating an inconsistency in speech output. Portions of text are missing in speech; shows that model is showing severe misalignment.
II. Tacotron 2 trained up to 200k steps with TinyMyST Dataset		Portions of the text are missing in speech at certain sections; which shows the model is skipping sections. There are possible errors or misinterpretations in the output.
III. Tacotron 2 pretrained with LJ speech for up to 100k steps and finetuned with TinyMyST for up to 200k steps		Text and speech show good alignment at the start but show a total disconnect between text and speech towards the end of the phrase; this indicates looping or stuttering in output.

The alignment plots derived from the Tacotron 2 training process are made available in Table 9. The training type I involved using the original MyST single-speaker dataset to train Tacotron 2. The plot from this training clearly shows that, even after 200k training

steps, there is no noticeable alignment. After that, the model was subjected to training with our cleaned version of the TinyMyST dataset (see Table 8). It seems that the model has a good grasp of the initial alignment but may struggle with longer sequences. This might be an indication that the model could benefit from additional training data or a review of its current training regime, particularly to ensure that it maintains strong alignment throughout longer sequences. Further experiments were conducted that incorporated transfer learning with a single-speaker model. The model was first trained with the LJ speech dataset and then finetuned with the TinyMyST. These experiments revealed certain patterns similar to those of child speech. An alignment plot of the generated speech from this experiment is illustrated in training type III in Table 9. The analysis of the plot indicates that while the phrases generated had a child-like pitch, the words and phonetic structures were unclear and indistinguishable. Additionally, the end of each phrase in the speech sample was characterized by meaningless white or static noise instead of distinct phonemes. This lack of clarity and coherence becomes more evident after listening to the generated synthetic speech. To improve further on this, the research progressed to exploring multispeaker TTS, aiming to synthesize varied child voices and enhance the synthetic speech dataset's diversity and realism.

4.1.2 MULTISPEAKER TTS TRAINING

While working on the multispeaker TTS research, we came across various multispeaker TTS models as mentioned earlier in Table 2. However, due to the lack of a well-established training framework for these approaches at the time of conducting this research, we decided against pursuing the training of these models using child speech. This decision stemmed from the challenges and uncertainties associated with implementing training protocols that were not yet fully defined or optimized. Instead, we focused our efforts on alternative methodologies that offered more structured and reliable frameworks such as the speaker verification-based approach, mentioned in the study [20]. Since this approach also utilized Tacotron 2 with more developed and proven training processes, it was employed to use this in our primary experiments. We build over this methodology for multispeaker TTS, however, we also include a two-stage training approach (pretraining and finetuning the acoustic model) for building more realistic child speech. We also use a different vocoder (WaveRNN) than used in their approach (WaveNet), which is used as a global vocoder for working with child speech (section 2.4.2). For the system to generate speech reflective of multiple speakers, it necessitates modifications to accommodate various speaker identities. Such enhancements have been realized in the multi-speaker TTS [20], which incorporates speaker embeddings to represent different speaker identities, feeding them into the Tacotron as an additional input. Consequently, the multispeaker TTS model [20] is composed of three distinct neural networks, each dedicated to a specialized task: the Speaker Encoder for speaker verification [131], the Acoustic model for synthesizing spectrograms [28], and the Vocoder [85] for creating the audio waveform. This structure is depicted in Figure 16 below.

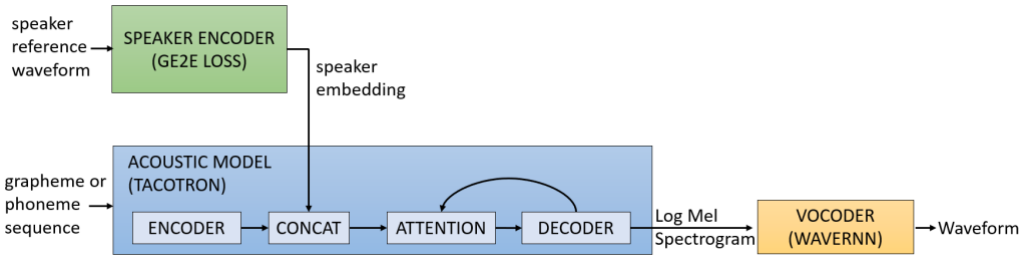


Figure 16: Multispeaker training pipeline.

The speaker encoder creates speaker embeddings by capturing the unique identity traits from the spoken utterances. Voices that are alike are positioned closer together in a latent space representation. Following this, the acoustic model produces spectrograms from text, while taking these speaker embeddings into consideration. These spectrograms are then transformed into audio waveforms by the vocoder. During the inference phase, a brief reference utterance (ground truth) of a child's voice is processed through the speaker encoder to generate the relevant speaker embeddings. The acoustic model then conditions its output on these embeddings. Initially, the acoustic model underwent training solely with adult speech data, particularly using the 'clean' dataset from Librispeech. This phase of training continued until convergence was observed at around 250k steps. Subsequently, the model was finetuned using the TinyMyST dataset, which contains child speech. This finetuning process was extended for an additional 750k steps. More comprehensive information about this process is available in our published paper in Appendix A.

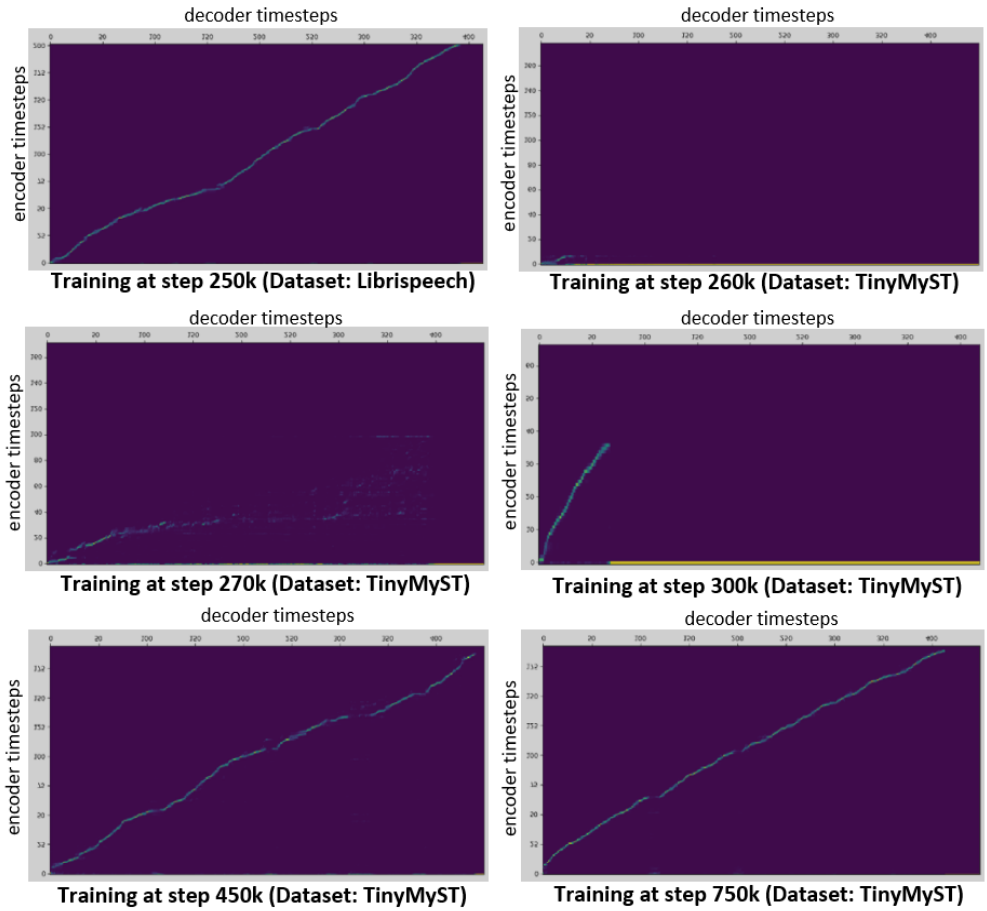


Figure 17: Alignment plots at different training steps for multispeaker Tacotron 2 involving transfer learning from adult to child speech.

Figure 17 displays the alignment plot for encoder-decoder timestamps. Initially, training on the LibriSpeech dataset for 250k steps yielded a strong alignment plot. The alignment initially weakened when switched to the TinyMyST dataset, improving as training progressed. The model underwent testing at various checkpoints, every 50k iterations, some of which are highlighted in Figure 17. Despite some alignments appearing similar, overall improvement in synthesized child voices was noted over time, assessed subjectively by listening to the output. The training was stopped at 750k steps, as no further significant improvements were observed after 700k steps, and the output waveform showed no additional enhancements. Some samples from this research work are presented online on

our GitHub page[1]. At this point, the vocoder is used to generate audio waveforms for performing the subjective evaluation.

4.1.3 SUBJECTIVE EVALUATION

Recognizing the novelty of our research in synthetic child speech, we determined that conventional MOS studies [132], [133], [134] would not suffice for assessment. Hence, we crafted a tailored MOS study, integrating aspects specifically relevant to child speech characteristics. To assess the phonetic range of our child's speech TTS, we utilized the Harvard sentences to evaluate the subjective quality of synthesized audio. Our evaluation employed a MOS-like method with various scoring categories. The evaluation process was streamlined through the use of a shared OneDrive environment, where all synthetic voices were accessible to evaluators, accompanied by a spreadsheet detailing the utterance IDs and transcript of the synthesized child's voice. This approach circumvented the challenges of understanding recorded child audio without transcripts, as some child speech can be unclear or nonsensical. Including transcripts allowed evaluators to correlate what they heard with what they read, enabling a more accurate assessment of the coherence between spoken phonemes and written graphemes.

Table 10: MOS (from 1 to 5) Explained for Speech Intelligibility, Voice Naturalness and Voice Consistency

Score	Speech Intelligibility	Voice Naturalness	Voice Consistency		
			Start of phrase	Middle of phrase	End of phrase
(5)	Clear	Consistent Tone	Clear Start	Clear Mid	Clear End
(4)	Mostly Clear	Minor Disjoint	Understandable	Minor Variation	Minor Distortion
(3)	Partial Clarity	Weak Consistency	Unclear Start	Notable Distort	Significant Noise
(2)	Muffled	Mixed Timbre	Distorted Start	Major Slurring	Unintelligible
(1)	Unintelligible	No Consistency	Incoherent	Incoherent	Incoherent

Evaluators were asked to rate the synthetic audio, which they listened to via headphones or earphones in a quiet setting on a scale of 1 to 5 across different categories in two phases which were developed over time. These categories included **Speech Intelligibility**, **Voice Naturalness,** and **Voice Consistency**, with the latter having three sub-categories: Start, Middle, and End of Phrase Quality. For this thesis, we have condensed the table explaining the MOS ratings (1 to 5) for each category in Table 10. However, the full version of this table, along with comprehensive details, can be found in Appendix A of our published work. This was also the first-ever MOS study conducted with synthetically generated child speech and we built this methodology for evaluation as a baseline for our future work. The results from this MOS study are presented in Table 11. A comparative analysis of the baseline MOS on Natural MyST and the synthetically generated utterances is also performed. There is some information loss observed at the end of most sentences containing inarticulate and unintelligent information or noise. The reason for this information loss can be redirected back to the child dataset used for training.

[1] https://c3imaging.github.io/ChildTTS/

Table 11: MOS Ratings Obtained From Subjective Evaluation With 95% Confidence Interval for Real and Synthetic Child Speech

	Speech Intelligibility	Voice Naturalness	Voice Consistency
Real Speech (from MyST)	4.21±0.42	4.05±0.34	4.08 ± 0.54
Synthetic Speech (Tacotron 2)	**3.95±0.30**	**3.89±0.32**	**3.96 ± 0.32**

4.1.4 OBJECTIVE EVALUATION

An objective evaluation methodology is also proposed for **Voice Naturalness, Speaker Similarity, and Speech Intelligibility**. We calculate objective Naturalness evaluation using a pretrained MOSNet, Speaker Similarity evaluation using a speaker verification system, and objective Intelligibility evaluation using a pretrained ASR system.

Voice Naturalness Evaluation Using a Pretrained MOSNet: MOSNet [133] predictions yield a high correlation to human ratings. As MOSNet was trained on adult speech, it is unlikely that it will generalize well for child speech. This objective evaluation was performed to see the correlation between reference child audio and synthetic child audio. The study revealed that the MOSNet score for reference child speech was 2.91, while synthetic child speech scored 2.60. The marginal MOS difference of 0.31 between these scores indicates a close correlation between real and synthetic child speech.

Speaker Similarity Evaluation Using a Speaker Verification System: Speaker similarity between synthesized and real speech is measurable using a Speaker Verification (SV) system [131]. The same pretrained speaker encoder from the TTS model training was employed alongside a tool named Resemblyzer[1] for extracting and visualizing speaker embeddings. Resemblyzer calculates similarity through cosine distance between embeddings. Additionally, we visualized this similarity in a 2D projection using the T-distributed stochastic neighbour embedding (T-SNE) [135], as shown in Figure 18. In this projection, 'gt' denotes ground truth child speech, while 'ss' represents labels for synthetic speech for the same child speakers. 'Adult_Male' and 'Adult_Female' labels correspond to randomly chosen male and female speakers from the Librispeech Dataset. In this 2D projection of speaker embeddings, Male and Female adult speakers are noticeably separated from each other and from child speakers.

[1] https://github.com/resemble-ai/Resemblyzer

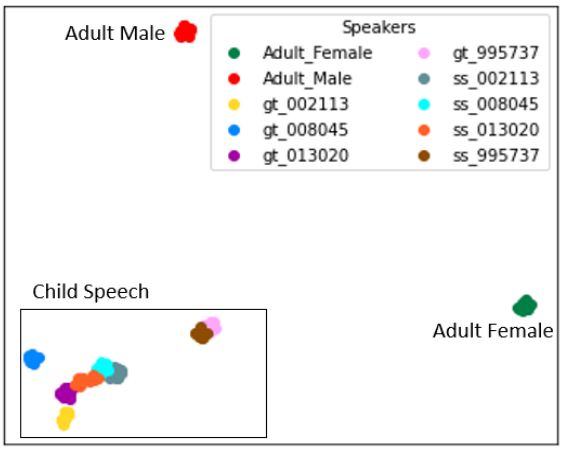


Figure 18: T-SNE 2D projections of speaker embedding for real child speech, synthetic child speech and adult speech. The child speech region, with real and synthetic child speech embeddings are marked inside a rectangle.

The cosine distance calculations between speaker embeddings revealed that the similarity between most adult and child speech lies in the range of 0.3-0.4. In contrast, the similarity between most synthetic and real child speech falls in the range of 0.65-0.85. With an average speaker similarity of 81% between synthetic and real child speech, it indicates that synthetic child speech closely resembles real child speech.

Objective Intelligibility evaluation using a pretrained ASR system: A pretrained[1] wav2vec2 model [17] evaluated synthetic speech intelligibility by comparing transcriptions of 120 real and synthetic child utterances. The model's performance was also assessed with adult speech from the LibriSpeech (test_clean) dataset, showing a WER of 3.43 for adult speech, reflecting the wav2vec2 model's training on such data. However, the WER rises to 15.27 for real child speech and further to 25.63 for synthetic utterances. The ASR model accurately recognized 75% of the synthetic speech, indicating a significant 10-point increase in WER compared to real child speech.

4.2 Multispeaker Fastpitch Methodology for Child Speech Synthesis

Fastpitch [21] integrates several unique features such as a duration predictor and self-alignment mechanism, which contribute to its advantages over Tacotron 2 [28] in certain aspects of TTS synthesis. Fastpitch is known for its faster inference speed, improved prosody control, and enhanced naturalness, which are crucial for capturing the dynamic range and expressiveness of child speech. The pitch prediction and duration prediction modules within Fastpitch provide more accurate control over the speaking rate and pitch, which are particularly variable in child speech compared to adults. These features justify Fastpitch as a superior approach for child speech synthesis, addressing some of the limitations that may be present in Tacotron 2's methodology, where pitch and prosody might not be as finely tuned. The motivation for this research is to overcome the challenges in generating synthetic child speech by providing more control over the prosody and duration of the generated speech. We also aim to overcome the challenges previously seen in the Tacotron 2 research. A transfer learning approach (Similar to section 4.1) to finetune a multi-speaker TTS model is applied with a cleaned version of the MyST dataset (referred

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md

to as MyST_2 in Chapter 3) using Fastpitch. An objective assessment, similar to that in section 4.1.6, has also been carried out to evaluate the naturalness, intelligibility, and speaker similarity of the speech generated by Fastpitch, and these findings are compared with those obtained using the Tacotron 2 method.

4.2.1 SINGLE SPEAKER TRAINING

In these experiments, we used the LJ Speech dataset [78] for finetuning with a single speaker. Initially, the model was trained on LJ Speech and then fine-tuned with one speaker from the MyST dataset. This approach resulted in quite noisy audio output. We also tried training on LJ Speech and then finetuning with the entire MyST dataset as if it were a single speaker, but this did not produce child-like speech. Consequently, we decided not to pursue single-speaker finetuning further.

4.2.2 MULTISPEAKER TRAINING

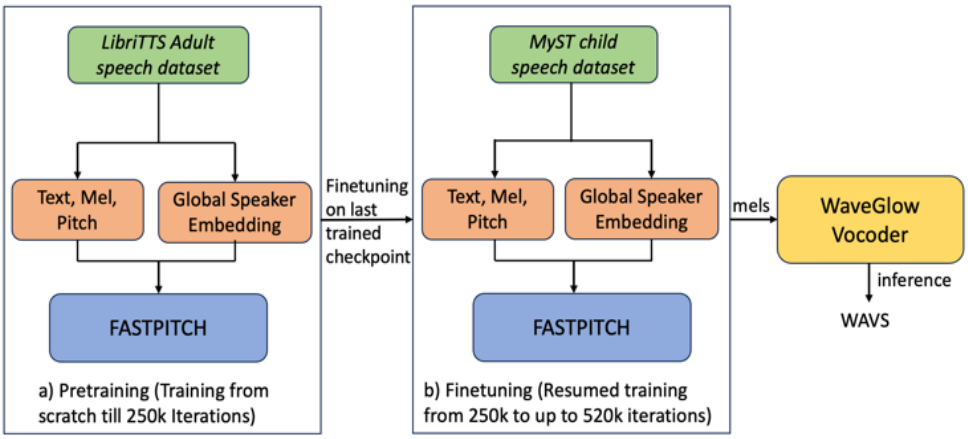
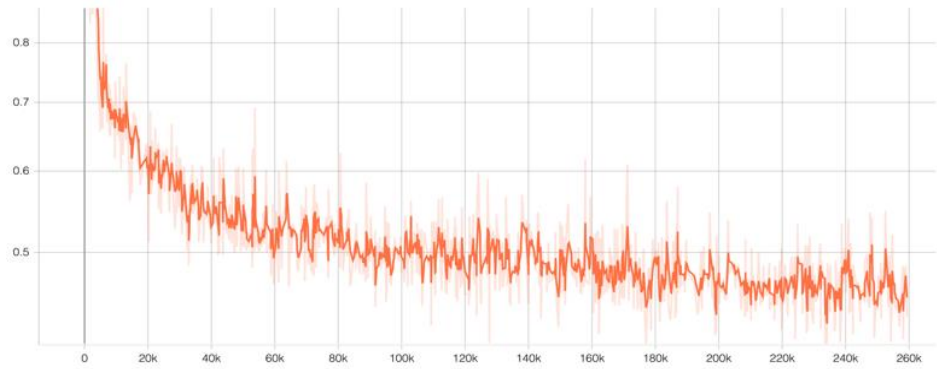
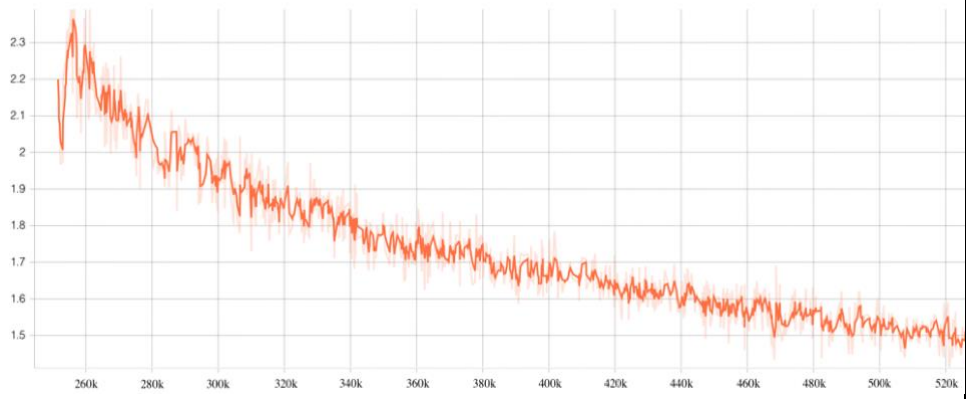


Figure 19: Transfer learning pipeline: a) Pretraining: Model being trained with LibriTTS dataset for up to 250k iterations. b) Finetuning: Resuming the acoustic model training with the MyST dataset from 250k iteration onwards up to 520k iterations.

The transfer learning pipeline was adopted from our previous approach using Tacotron 2 [29]. Figure 19 describes the transfer learning pipeline adopted for Fastpitch [21]. Fastpitch uses the global speaker embeddings [136], which means that it learns a distinct embedding for each speaker during training. These embeddings are learned along with the other parameters of the model. This is in contrast to some other systems where speaker embeddings might be generated by a separate, external speaker encoder (as seen in some implementations of Tacotron 2). By learning speaker embeddings within the model itself, Fastpitch can more seamlessly integrate the characteristics of different speakers into its speech generation process. This internal approach can lead to more coherent and natural-sounding speech synthesis, especially when the model is trained with a diverse set of speakers. In this work, the model is first trained with the LibriTTS dataset (585 hours) for up to 250k iterations until a consistent low loss threshold is achieved, and the model starts to converge. After that, the model was finetuned from 250k iterations onwards for up to 520k additional steps using the MyST_train dataset (55 hours). These datasets were earlier described in Table 8.

Table 12: Loss Curves for Multispeaker Fastpitch Training

Loss Curve	Observations
 **I. Initial Training:** Training with LibriTTS up to 250k iterations.	Loss decreased steadily and stabilized at 250k steps.
 **II. Finetuning:** Resume training on MyST_train from 250k up to 520k iterations.	Loss initially increased, then declined and plateaued; overfitting was observed post 550k steps.

In our multispeaker TTS training experiments, the model initially trained on the LibriTTS dataset showed a loss curve depicted in Table 12.I. The loss decreased gradually for the first 2000 warmup steps, then steadily until stabilizing at an average of 0.3 around the 250k steps mark. With no further loss improvements, training was paused for finetuning. Finetuning then continued on the MyST_train dataset from the 250k steps up to 520k steps (see Table 12.II), where loss initially increased before starting to decline around the 260k epoch, eventually plateauing at the 520k step with no significant further improvement. This was confirmed by listening to the audio files generated at every 50k epoch interval. Post 550k epochs, the model showed signs of overfitting, learning noise features from the MyST dataset, which degraded the audio quality. Using this methodology, we have also released a Synthetic child speech dataset generated from this research, which will be discussed in more detail in Chapter 5.

4.2.3 OBJECTIVE EVALUATION

We conducted objective evaluations, specifically focusing on the aspects of Naturalness, Intelligibility, and Speaker similarity (as previously done in section 4.1.4) to provide a comparative analysis with our previous work involving child speech synthesis using the Tacotron 2 model.

Objective Naturalness Evaluation Using the Pretrained MOSNet [133]: Table 13 displays the MOS for speech samples assessed using the pretrained MOSNet model. The Fastpitch model's synthetic speech surpassed both the original and Tacotron 2-generated versions in quality, showing a strong correlation between synthetic child speech and real child speech.

Table 13: Objective Evaluation Using Pretrained MOSNet for Fastpitch

Dataset	MOS
Adult speech (Librispeech test_clean)	3.78 ± 0.07
Original Child Speech [MyST]	2.91 ± 0.07
Tacotron 2-based synthetic child speech [29]	2.60 ± 0.06
Fastpitch-based synthetic child speech [21]	**3.10 ± 0.12**

Objective Intelligibility Evaluation Using a Pretrained wav2vec2 ASR System: The wav2vec2 base model[1] was used to measure the objective intelligibility of the Fastpitch-generated synthetic child speech. Table 14 details the WER across various speech datasets. Our implementation using the Fastpitch model yielded a WER of 17.61, closely aligning with the WER of original child speech from the MyST dataset. This performance also exceeded that of Tacotron 2-generated child speech.

Table 14: Objective Intelligibility Evaluation Using a Pretrained ASR for Fastpitch

Dataset	WER
Adult Speech (Librispeech test_clean)	3.43
Original Child Speech [MyST]	15.27
Tacotron-2 based synthetic child speech [29]	25.63
Fastpitch-based synthetic child speech [21]	**17.61**

Speaker Similarity Verification Using a Pretrained Speaker Verification System: Similar to Tacotron 2, the pretrained speaker encoder [131] from Resemblyzer[2] was used to extract and visualize the speaker embeddings. The 2D plot for Fastpitch speaker embedding visualization (Figure 20) looks very similar to that of Tacotron 2 (Figure 18). This clustering analysis indicates a correlation among child speakers, encompassing both real and synthesized speech, in contrast to the more dispersed adult male and female speakers.

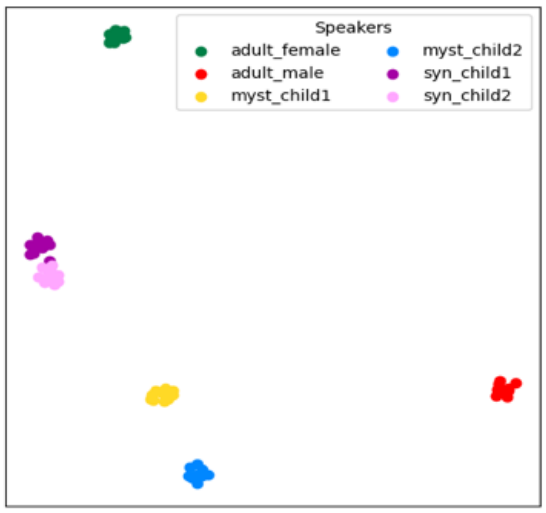


Figure 20: T-SNE embedding projections for actual child speech ('myst_child1'-boy and 'myst_child2'-girl), Fastpitch-generated synthetic child speech ('syn_child1' and 'syn_child2'), and adult speech ('adult_female' and 'adult_male') speakers.

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md
[2] https://github.com/resemble-ai/Resemblyzer

To quantify the similarity between real child speech and synthetic child speech, we used cosine similarity measures. These measures showed that child-to-adult speech similarities fall between 0.34 to 0.53, while synthetic child speech closely matches real child speech, with similarities ranging from 0.63 to 0.98. The synthetic and real child voices share an average similarity of 77%. This percentage is a bit lower than the 81% achieved for the Tacotron 2 similarity measure. We believe this could be due to an increase in the training dataset from 19 hours in Tacotron 2 to 55 hours in Fastpitch. This increase in data also encompasses more speakers and increases the complexity of acoustic model training.

4.3 Conclusion and Final Remarks

Our primary contribution with Tacotron 2 showcases a significant advancement in the synthesis of child speech with limited data, presenting a viable solution to the challenge of synthesizing diverse and natural-sounding child voices. The research not only furthers the capabilities of TTS models but also opens pathways for creating large synthetic datasets that could serve various applications within the field of child speech research, such as ASR and speaker recognition. Our subjective assessment reveals a noticeable loss of information towards the end of the generated speech. This is also reflected in a lower MOS for the 'End of Phrase' category under Voice Consistency. We also demonstrate the successful application of the Fastpitch TTS model in synthesizing child voices with limited training data. This is a significant advancement in the field of synthetic child speech generation. Additionally, we release a synthetic dataset generated from this work which will be discussed in more detail in Chapter 6. We acknowledge the lack of subjective evaluation in assessing the 'Naturalness', 'Intelligibility', and 'Speaker Similarity' of the generated child speech using Fastpitch. The decision to forgo a subjective MOS evaluation was primarily driven by time constraints. While objective measures like MOSNet provide valuable insights, they may not fully capture the nuances of human perception.

While the MyST dataset provides a basis for training, its scope and diversity could limit the model's ability to generalize across various child speech characteristics. Therefore, we believe that the creation of more diverse and extensive child speech datasets is essential for future progress. With this in mind, we decided to delve into ASR research for child speech. The aim is to train state-of-the-art ASR models for improved transcription of the untranscribed MyST dataset (197 hours) and other child speech datasets from Xperi data acquisition. A detailed experimental study for both Tacotron 2 and Fastpitch methodologies for child speech synthesis along with the proposed evaluations are explained in detail in our published paper attached in Appendix A and Appendix D.

Chapter 5

Contribution to Improving ASR Technologies for Child Speech

In the initial stages of the DAVID project, the emphasis was on developing TTS technologies for Edge-AI applications. However, approximately a year into the research, it was observed that the available public datasets for child speech were of poor quality. This revelation led the project towards investigating ASR techniques as a strategy to clean up public child-speech datasets which were of poor quality. We aimed to use ASR technologies to employ it as a transcription tool for providing better annotations and creating larger child speech datasets. However, it was soon discovered that ASR models, predominantly trained on adult speech, were poor at transcribing child speech data effectively. It was evident that there was a need to research state-of-the-art neural ASR models to better accommodate child speech. Given that our industry partners lacked expertise in the latest neural models that had emerged since the start of the DAVID project, a comprehensive evaluation of these models' adaptability to child speech became necessary. The objective set by one industry partner to gather child speech data for training Edge-AI models was also impeded by the COVID pandemic, leading to considerable delays. Consequently, the project's direction shifted towards the development of generative child-TTS technologies. Therefore, investigating the potential of child-speech ASR became important to enhance the DAVID project's capabilities in assessing the quality of generated TTS and annotating newly collected child speech data by the industry partner.

In initiating our exploration of ASR systems, our review extended various methodologies aimed at optimizing ASR for children's speech [3], [4]. Among the diverse methodologies, several stood out due to their effectiveness and adaptability to the nuanced characteristics of child speech. These methodologies encompass Transfer Learning methods [11], [137], [138], [139], Hidden Markov Models (HMMs)-based methods [52], [140], [141], Augmentation-based methods [107], [142], [143], [144], [145], Self-Supervised learning (SSL) [146], [147], [148], [149], [150], and semi-supervised approaches [18], [149], [151]. Given the distinct challenges associated with child ASR, including data scarcity and the inherent variability in child speech, Transfer Learning emerged as a particularly effective method. This approach leverages the abundance of adult speech data to pretrain models, which are subsequently finetuned on child speech datasets. Such a methodology not only addresses the issue of limited child speech data but also allows the model to initially learn from the acoustic features present in adult speech. Consequently, our experimental framework is designed around a series of experiments that involve pretraining on adult speech followed by finetuning on child speech.

In this chapter, we address the challenges in child speech recognition by using the novel approach of wav2vec2 [17]. An in-depth experimental analysis is performed including pretraining and finetuning on different seen and unseen child speech datasets. We also clean and prepare datasets for child speech ASR (which is also made available for research

use). Finally, we provide a comparative analysis with the previous SOTA research on a similar distribution of datasets. Additionally, the evaluation of Whisper models [18] for child speech recognition is presented, showcasing their adaptability to the distinct nuances of child speech. Significant enhancements in ASR performance were observed when Whisper models were finetuned on child speech datasets, surpassing results from models without finetuning. This was further evidenced by testing Whisper models [18] with various non-native English child speech accents, including American, British, German, Italian, Swedish, and Chinese. Subsequent finetuning experiments were also performed over a combination of non-native child speech datasets. We report substantial improvements in performance on various non-native English child speech datasets compared to previous state-of-the-art results. Lastly, we adapted Conformer-Transducer models [16] for the task of child speech recognition in a similar way to Whisper and wav2vec2. The key was the comparative analysis of these models, all finetuned on the same child speech data, which could provide valuable insights into the efficacy of these models for child speech recognition.

5.1 Self-Supervised Learning Approach Using wav2vec2 ASR

This section details a series of experiments and findings focused on using the SSL-based wav2vec2 model [17], to enhance ASR for child speech. We use a combination of adult and child speech datasets for pretraining and finetuning to find the best experimental configuration to work with child speech. The pretraining was conducted on a mix of unlabelled child and adult speech datasets. It's also worth noting that wav2vec2's pretraining architecture transitions from a 'base' to a 'large' configuration typically involving scaling up model parameters to handle more extensive datasets effectively. The 'base' model is designed for efficiency and quick training on smaller datasets (such as MyST_complete and Librispeech), whereas the 'large' version, with a greater number of layers, is designed to capture intricate patterns in larger volumes of data (such as Librilight). For finetuning, subsets of varying lengths were created from the MyST (MyST_10m, MyST_1h, MyST_10h, MyST_55h*)* and PFSTAR (PFS_10m, PFS_1h, PFS_10h) datasets to determine the optimal data requirement for child speech experiments. These ranged from 10-minute to 55-hour segments, formed by randomly selecting files from larger sets (MyST_55h and PFS_10h). Finally, the evaluation of ASR was based on the WER across various child speech datasets, MyST_test (10hrs), PFS_test (2hrs), and CMU_Kids (9hrs), and adult speech LibriTTS dev-clean dataset (8.9 hrs). Experiments were categorized into five groups: A, B, C, D, and E. These experiments are grouped in Table 15, encompassing 32 experiments with various pretraining and finetuning dataset combinations. The goal was to explore cross-domain WER correlations across pretraining and finetuning datasets. Notably, no models were pretrained solely on child speech due to the lack of ample publicly available data for effective learning of child speech representations.

Table 15: WER Obtained for wav2vec2 Finetuning Experiments Over MyST, PFSTAR and CMU_Kids Datasets

	Groups	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMU_KIDS	WER dev_clean
1	A	Librispeech	LS_10m	31.48	30.05	33.38	15.90
2			LS_100h	17.82	15.96	18.73	4.16
3			LS_960h	**15.41**	**11.20**	**16.33**	**3.40**
4			LS_10m	26.47	27.14	29.37	15.35

5		Librilight	LS_100h	13.15	11.63	16.18	3.79
6			LS_960h	**12.50**	**8.56**	**14.85**	**3.28**
7	B	Librispeech	MyST_10m	28.84	41.34	34.18	21.45
8			MyST_1h	18.75	31.84	23.13	13.91
9			MyST_10h	13.46	28.68	19.59	10.94
10			MyST_55h	**8.13**	**14.77**	**16.47**	**7.72**
11		Librilight	MyST_10m	33.01	44.36	39.91	46.45
12			MyST_1h	14.91	26.21	18.74	11.59
13			MyST_10h	12.92	25.05	17.72	10.04
14			MyST_55h	**7.51**	**12.46**	**15.25**	**6.43**
15	C	Librispeech MyST_ -complete	MyST_10m	29.16	45.71	37.56	35.39
16			MyST_1h	21.89	38.53	29.03	20.45
17			MyST_10h	16.18	32.95	25.06	16.83
18			MyST_55h	**10.34**	**25.47**	**23.15**	**13.48**
19	D	Librispeech	PFS_10m	35.91	16.43	33.53	30.43
20			PFS_1h	33.52	7.36	29.55	16.61
21			PFS_10h	**31.86**	**3.48**	**27.49**	**13.95**
22		Librilight	PFS_10m	37.10	16.78	35.13	23.85
23			PFS_1h	30.81	14.19	28.54	21.89
24			PFS_10h	**27.17**	**3.50**	**21.35**	**11.60**
25	E	Librispeech	LS_960h, MyST_55h	8.18	12.17	14.12	1.24
26			LS_960h, PFS_10h	15.42	3.74	15.31	1.41
27			MyST_55h, PFS_10h	**7.94**	**2.91**	15.97	7.64
28			LS_960h, MyST_55h, PFS_10h	8.13	3.12	**13.76**	**1.20**
29		Librilight	LS_960h, MyST_55h	8.06	9.31	13.20	1.34
30			LS_960h, PFS_10h	13.18	3.17	13.19	**1.32**
31			MyST_55h, PFS_10h	**7.42**	**2.99**	14.18	5.79
32			LS_960h, MyST_55h, PFS_10h	8.17	3.33	**12.77**	1.40

Group-A: The pretrained BASE and LARGE models from wav2vec2, trained with 960 hours of Librispeech and 60k hours of Librilight dataset were finetuned with 10 minutes, 100 hours, and 960 hours of Librispeech. The objective was to observe the performance of child speech without any being used in training. Models showed a decrease in WER with an increase in finetuning dataset size. Interestingly, only a small difference in WER was observed between the BASE and LARGE models, despite the LARGE model having 60 times more training data. We also observed Significant finetuning improvements from 10 minutes to 100 hours of adult speech, with diminishing returns beyond 100 hours.

Group-B: This group used similar pretraining configurations as Group-A. The finetuning used 10 minutes, 1 hour, 10 hours, and 55 hours of MyST child speech data. The aim was to introduce a single child speech dataset to the model training and observe its impact during inference on child speech. Lower WERs were achieved for child speech in comparison to Group A. A trend of decreasing WER with the increasing finetuning data was noted. We also observed that as little as 1 hour of child speech data showed improvements comparable to 100 hours of adult speech data and similarly, 10 hours of child speech matches 960 hours of adult improvements. Domain mismatch was also observed with weaker improvements on the PFS_test and CMU_Kids compared to the MyST_test.

Group-C: This group involved pretraining with Librispeech and MyST_complete. The goal was to integrate child speech data during the pretraining phase of the wav2vec2 model along with adult speech. Finetuning employed the same MyST data volumes as in Group B. This approach resulted in increased WERs compared to BASE models from Group B, indicating that adding child speech in pretraining was not effective. As a result, finetuning with Large configurations was not conducted.

Group-D: This group mirrored Group-A's BASE and LARGE adult speech pretraining but excluded MyST_Complete due to its lack of impact in group-C. For finetuning, the PFSTAR dataset, segmented into 10 minutes, 1 hour, and 10 hours, replaced MyST. While the PFS_test showed lower WERs, increases in WER were observed on other unseen datasets highlighting domain mismatches.

Group-E: Pretraining used BASE and LARGE configuration as group-A and experimented with mixed finetuning datasets including LS_960h+MyST_55h, LS_960h+PFS_10h, MyST_55h+PFS_10h, and LS_960h+MyST_55h+PFS_10h combinations. The aim was to test various training combinations on test datasets comprising child speech. This group showed the best WERs across all datasets, indicating the importance of domain match in finetuning data. Both BASE and LARGE configurations showed similar performance with cross-domain dataset finetuning.

To conclude, the pretraining models with adult speech data alone outperform those using a mix of adult and child speech, especially when incorporating low-quality datasets like MyST, which diminishes ASR model performance across all tests. Domain differences are significant, with PFSTAR showing higher quality than MyST and CMU_Kids, though the latter two align more closely. Cross-domain child speech finetuning yields the best outcomes. The wav2vec2 BASE configuration, needing far less data than the LARGE variant, suits low-data scenarios well, with the LARGE's slight improvements not compensating for its higher computational cost. Finetuning with 100 hours of adult speech strikes a balance between effort and accuracy, but even 10 hours of child speech data finetuning surpasses using extensive adult speech data, with around 65 hours of mixed-domain child speech data proving optimal.

5.1.1 COMPARISON WITH PREVIOUS SOTA APPROACHES

Previous researchers have employed different methods to clean and use the data, but the lack of a standardized approach makes direct comparisons challenging. However, our research using the wav2vec2 approach has shown promising results, significantly improving over previous studies on the same datasets as detailed in Table 16. Additionally, our method of 'cleaning' the test datasets is thoroughly explained earlier in Chapter 2, providing a foundation for future research comparisons.

Table 16: Comparison Between Previously Obtained SOTA Results and Our Results on the MyST, PFSTAR and CMU_Kids Dataset

SOTA Papers	Method Type	WER MyST	WER PFSTAR	WER CMU_Kids
TDNN-F + Augmentation [152]	Supervised	-	-	16.01
Hybrid HMM-DNN Transfer Learning [138]	Supervised	-	-	19.33
DRAFT [148]: ▪ wav2vec2 ▪ HuBERT	SSL	 16.70 16.53	-	-
Transformer + CTC + Greedy [153]	Supervised	16.01	-	-
W2V2 + source-filter warping + LM [139]	SSL		4.86	
Our Work	**SSL**	**7.42**	**2.91**	**12.77**

Our work provided a baseline for future experiments. The experimental framework described for finetuning the wav2vec2 model was also utilized to train the Whisper and Conformer-Transducer models to provide a comparative analysis (will be discussed in more detail in consecutive sections). We also use this experimental framework as an objective evaluation method for validating speech augmentation experiments which will be discussed in section 6.2. The complete working methodology along with the detailed experimental results for the wav2vec2 experiments is presented and published in IEEE Access titled "A wav2vec2-based Experimental study on Self-Supervised Learning Methods to Improve Child Speech Recognition" with a copy included in Appendix B of this thesis.

5.2 Supervised Learning Approach Using Whisper ASR

This is a follow-up work based on the experimental framework proposed with wav2vec2 experiments. The original Whisper models [18] were initially evaluated on different child speech datasets, including MyST_test, PFS_test, and CMU_test without any initial finetuning. The models were categorized by size (Tiny, Base, Small, Medium, Large, and Large V2) and further divided into two language training versions: a multilingual one and an English-only version (denoted by '.en' in the name). Including English models also allows for comparing the performance of same-language training with child speech datasets against multilingual models with non-English datasets. The results from this experiment can be seen in Table 17.

Table 17: WER Obtained for Whisper Models (No-Finetuning)

ID	Models	MyST_ test	PFS_ test	CMU_ test	dev-clean
1	Tiny	40.09	159.57	30.63	10.85
2	Tiny.en	33.02	47.11	27.32	8.62
3	Base	32.14	100.07	25.03	8.14
4	Base.en	29.15	45.70	20.75	7.18
5	Small	26.22	111.75	18.52	6.43
6	Small.en	26.72	39.00	16.82	6.06
7	Medium	**25.11**	80.97	**12.67**	5.58
8	Medium.en	28.06	**35.25**	14.00	6.20
9	Large	**25.24**	84.52	13.70	5.53

10	Large-V2	**25.00**	73.68	**12.69**	**5.40**
11	w2v2-base (Table 15.3)	**15.41**	**11.20**	**16.33**	**3.40**
12	w2v2-large (Table 15.6)	**12.50**	**8.56**	**14.85**	**3.28**

Note: w2v2-base (ID_11) and w2v2-large (ID_12) are equivalent to model 3 and model 6 from Table 15, which are shown here to provide a comparison with whisper models.

These Whisper models (without finetuning) showed varying performance on the child speech datasets. Larger models generally performed better, with English-only models outperforming multilingual ones in language-specific tasks. Despite their effectiveness in adult speech recognition, these models performed poorly on child speech without finetuning. In contrast**,** wav2vec2 models (with no child speech finetuning) generally performed better on child speech compared to non-finetuned Whisper models. The best-performing models (i.e. Medium, Medium.en and Large-v2) from this initial evaluation were selected for further finetuning with child speech datasets. Three subsets of experiments involved finetuning with MyST_55h, PFSTAR_10h, and a combination of both datasets. The finetuned Whisper models were compared with wav2vec2 models that had also been finetuned on similar distribution of child speech datasets as can be seen in Table 18 which provides the WER for the finetuning experiments.

Table 18: WER obtained for Finetuning Whisper and wav2vec2 Models With Child Speech Datasets

ID	Models	MyST_ test	PFS_ test	CMU_ test	dev-clean
MyST (55 Hours) Finetuning:					
1	Medium	**11.66**	19.76	16.84	5.62
2	Medium.en	11.81	17.83	**15.07**	6.48
3	Large-V2	12.28	**10.88**	15.67	**4.82**
4	w2v2-base (Table 15.10)	8.13	14.77	16.47	7.72
5	w2v2-large (Table 15.14)	**7.51**	**12.46**	**15.25**	**6.43**
PFSTAR (10 Hours) Finetuning:					
6	Medium	16.18	3.15	16.57	5.33
7	Medium.en	15.84	3.14	15.53	5.28
8	Large-V2	**15.79**	**2.88**	**15.22**	**5.10**
9	w2v2-base (Table 15.21)	31.86	**3.48**	27.49	13.95
10	w2v2-large (Table 15.24)	**27.17**	3.50	**21.35**	**11.60**
MyST (55 Hours) + PFSTAR (10 Hours) Finetuning:					
11	Medium	**12.22**	**2.98**	16.05	5.40
12	Medium.en	12.33	3.32	**15.08**	**4.88**
13	Large-V2	13.34	4.17	17.11	4.97
14	w2v2-base (Table 15.27)	7.94	**2.91**	15.97	7.64
15	w2v2-large (Table 15.31)	**7.42**	2.99	**14.18**	**5.79**

Note: Results for wav2vec2 models with IDs 4, 5, 9, 10, 14, and 15 are taken from Table 15, which are presented here for comparison with Whisper models.

Finetuning Whisper models on child speech datasets notably enhanced their accuracy, with MyST_train finetuning yielding significant WER improvements, particularly on the MyST_test and PFS_test datasets. Similarly, PFSTAR_train finetuning boosted WER on the PFSTAR_test dataset, although it had a lesser impact on other datasets. Utilizing both MyST and PFSTAR for finetuning enhanced model performance on datasets of similar distributions while maintaining strong WER on unseen datasets. Overall, Whisper models

improve ASR across adult and child speech, showing versatility across various finetuning datasets and unseen datasets. In contrast, wav2vec2 excels with finetuning-specific datasets, offering superior performance in task-specific applications. The wav2vec2 finetuning achieved lower WER on the MyST_test dataset and, when finetuned with a mix of child speech datasets, exhibited good performance across all child speech datasets. Furthermore, wav2vec2's smaller model size and its requirement for ten times less training data than Whisper make it an ideal choice for deployment on edge devices. This distinction highlights Whisper's broad applicability and wav2vec2's optimization for specific tasks, underlining the strategic choice between the two depending on the application context and deployment constraints. The complete working methodology along with the detailed experimental results and comprehensive discussion of this topic is presented and published in the Interspeech 2023 Conference titled "Adaptation of Whisper Models to Child Speech Recognition", a copy of which is made available in Appendix C of this report.

5.3 Comparison Between wav2vec2, Whisper and Conformer Models

The Conformer [16] model's codebase and training were undertaken by a fellow PhD student, Andrei Barcovschi, while my contribution focused on designing experiments, preparing datasets, and providing a comparative analysis with the wav2vec2 and Whisper approaches. This work represents a continuation of our previous efforts in dataset preparation, enabling comprehensive comparisons among the three approaches. We first evaluate the original conformer transducer models provided by Nvidia[1] on different child speech datasets without any finetuning involved and later these models were finetuned on different combinations of child speech datasets (MyST and PFSTAR) to provide a comparison with Whisper [18] and wav2vec2 [17] approaches.

Table 19: Comparison Between Conformer, Whisper and wav2vec2 Models (Without Any Finetuning on Child Speech)

ID	ASRs	Models	MyST_test	PFS_test	CMU_test
1	Conformer-Transducer	Small	**21.34**	12.68	16.05
2		Medium	24.99	11.58	17.51
3		Large	25.91	8.94	15.06
4		Xlarge	24.42	**8.22**	**14.83**
5	Whisper	Small.en (Table 17.6)	26.72	39.00	16.82
6		Medium (Table 17.7)	25.11	80.97	**12.67**
7		Medium.en (Table 17.8)	28.06	**35.25**	14.00
8		Large-V2 (Table 17.10)	**25.00**	73.68	12.69
9	wav2vec2	w2v2-base (Table 15.3)	15.41	11.20	16.33
10		w2v2-large (Table 15.6)	**12.50**	**8.56**	**14.85**

Note: Whisper and wav2vec2 results were previously presented in Table 15 and Table 17. They are made available here for comparison with Conformer models.

Table 19 displays the WERs of the original Whisper, wav2vec2, and Conformer-transducer models on child speech datasets without any initial finetuning. There's a noticeable trend of high WERs, around 25%, in the MyST_test set for most Whisper and Conformer-

[1] https://github.com/NVIDIA/NeMo/blob/main/nemo/collections/asr/README.md

transducer models, except for wav2vec2 models, which show about 10% lower WERs. Smaller Conformer-transducer models outperform their Whisper counterparts, especially on child speech datasets. However, larger Conformer-transducer models lose this edge, while larger Whisper and wav2vec2 models demonstrate better performance, indicating a potential loss of generalization in larger Conformer-transducer models. The 'Large' and 'Xlarge' Conformer-transducer models show competitive results in most cases. Whisper models have generally higher WERs, except for the 'Medium' and 'Large' models which perform well on all datasets. The 'w2v2-large' model stands out with the lowest WERs among all evaluated models.

Table 20: WER for Different Whisper, wav2vec2 and Conformer Models Finetuned on MyST, PFSTAR and a Combination of Both Datasets

Name	Models	MyST_test	PFS_test	CMU_test
MyST (55 Hours) Finetuning:				
Conformer-Transducer	Large	14.17	44.02	27.03
	XLarge	13.79	43.57	20.63
Whisper	Medium.en (Table 18.2)	**11.81**	17.83	**15.07**
	Large-V2 (Table 18.3)	12.28	**10.88**	15.67
wav2vec2	w2v2-base (Table 15.10)	8.13	14.77	16.47
	w2v2-large (Table 15.14)	**7.51**	**12.46**	**15.25**
PFSTAR (10 Hours) Finetuning:				
Conformer-Transducer	Large	90.00	8.58	82.00
	XLarge	86.79	**6.31**	75.26
Whisper	Medium.en (Table 18.7)	15.84	3.14	15.53
	Large-V2 (Table 18.8)	**15.79**	**2.88**	15.22
wav2vec2	w2v2-base (Table 15.21)	31.86	**3.48**	27.49
	w2v2-large (Table 15.24)	**27.17**	3.50	**21.35**
MyST (55 Hours) + PFSTAR (10 Hours) Finetuning:				
Conformer-Transducer	Large	13.86	4.44	25.00
	XLarge	13.61	4.30	21.21
Whisper	Medium.en (Table 18.12)	12.33	3.32	15.08
	Large-V2 (Table 18.13)	13.34	4.17	17.11
wav2vec2	w2v2-base (Table 15.27)	7.94	**2.91**	15.97
	w2v2-large (Table 15.31)	**7.42**	2.99	**14.18**

Note: Whisper and wav2vec2 results were previously presented in Table 15 and Table 18. They are made available here for comparison with Conformer models.

Finetuning experiments on Conformer-transducer models used Large and Xlarge versions, on Whisper models included Medium.en and Large-V2, while wav2vec2 involved its base and large versions. Table 20 compares their WERs on evaluation sets. Conformer-transducer models showed increased WERs on the PFS_test and CMU_test when finetuned with MyST_55h, indicating poorer handling of noisy datasets compared to Whisper and wav2vec2 models. Finetuning the Conformer-transducer on PFS_10h reduced WER on the PFS_test but still underperformed compared to Whisper and wav2vec2. Combined dataset finetuning improved the Conformer-transducer's performance across all datasets, suggesting better generalization with diverse training. However, the benefits of larger models are marginal considering their higher computational demands. Models finetuned on MyST perform better on the MyST_test, and those on PFSTAR show improvements on

the PFS_test. This highlights the importance of domain-specific finetuning. Whisper provides good generalization to unseen datasets without losing any previous information after finetuning. Overall, wav2vec2 models, being smaller and requiring less data, appear to be the most efficient for child speech ASR, outperforming others across various datasets and achieving the lowest WERs.

The complete working methodology along with the detailed experimental analysis is presented and published in the Sped 23 conference titled “A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition” [19]. A copy of the published paper based on this section is attached and presented in Appendix F of this report.

5.4 Whisper Approach to Improving ASR for Non-Native Child Speech

This is a follow-up work on the Whisper Models [18]. The primary contribution lies in adapting and finetuning the Whisper for non-native child speech data. We use the datasets presented in Table 8 from different non-native English child speech including American-accented English, British-accented English, Swedish-accented English, German-accented English, Italian-accented English and Chinese-accented English.

Dataset Preparation: MyST_train and MyST_test are the same as mentioned in Table 8. PFSTAR British dataset was divided into 10 hours for training, called ‘PF_br_train’, and 2 hours for testing, called ‘PF_br_test’. Similarly, the rest of the non-native PFSTAR dataset was divided into an 8:2 ratio for testing and training. The intention here was to have more data during inference and use a low amount of non-native datasets for finetuning (training). Therefore, the PFSTAR Swedish subset was divided into 1.01 hours for testing (PF_sw_test) and 0.25 hours for training (PF_sw_train). Similarly, the PFSTAR German subset was divided into 2.55 hours for testing (PF_ge_test) and 0.68 hours for training (PF_ge_train). The PFSTAR Italian subset was divided into 2.8 hours for testing (PF_it_test) and 0.7 hours for training (PF_it_train). The Speechocean762 containing Chinese-accented English was divided into 1.92 hours for testing (SO_test) and 0.48 hours for training (SO_train). Since the amount of non-native data is small, we combined all the non-native training datasets mentioned in the previous section into two subsets used for finetuning: Non_Native_10 (NN_10), which contains half of the non-native training sets, comprising half of PF_sw_train, PF_ge_train, PF_it_train, and SO_train, corresponding to 10% of the total non-native training data available and Non_Native_20 (NN_20), which includes complete set of PF_sw_train, PF_ge_train, PF_it_train, and SO_train, amounting to 20% of the total non-native training datasets available.

A series of experiments were conducted categorized into nine Groups (A to I) to evaluate the performance of Whisper models on various child speech datasets. In Group-A, Original Whisper models were tested on different child speech datasets without any finetuning. This set of experiments served as the baseline to assess the model's inherent capabilities. The top-performing models from Group-A were selected for further finetuning in Groups B to I. Different combinations of child speech datasets were used for finetuning to determine the optimal mix for lowering WERs. The focus was on adapting Whisper models to specific child speech datasets, particularly examining the impact of non-native English datasets. The results obtained from these experiments as well as information regarding the finetuning groups can be found in Table 21.

Table 21: WER Obtained for Different Group Experiments With Whisper Models

ID	Models	MyST_test	PF_br_test	CMU_test	PF_sw_test	PF_ge_test	PF_it_test	SO_test	Dev_clean
Group A: No-Finetuning:									
1	Tiny	40.09	159.57	24.62	55.32	103.68	70.57	64.83	10.85
2	Tiny.en	33.02	47.11	16.25	45.23	89.80	47.22	51.28	8.62
3	Base	32.14	100.07	16.65	53.88	126.84	50.29	60.39	8.14
4	Base.en	29.15	45.70	15.01	37.29	93.77	46.84	38.47	7.18
5	Small	26.22	111.75	9.30	60.81	86.72	44.09	36.19	6.43
6	Small.en	26.72	39.00	8.64	32.26	**71.04**	33.38	30.33	6.06
7	**Medium**	**25.11**	80.97	**7.48**	35.07	105.82	45.65	37.00	**5.58**
8	**Medium.en**	28.06	**35.25**	**7.17**	**27.91**	80.40	**25.94**	**25.29**	6.20
9	Large	25.24	84.52	7.56	33.09	79.14	51.82	37.25	5.53
10	**Large-V2**	**25.00**	73.68	**6.86**	**29.99**	**77.56**	**34.97**	**29.39**	**5.40**
Group B: MyST_train Finetuning:									
11	Medium	**11.66**	19.76	9.43	34.18	**62.40**	24.53	24.89	5.62
12	Medium.en	11.81	17.83	**9.13**	**23.63**	76.84	**19.99**	25.45	6.48
13	Large-V2	12.28	**10.88**	9.80	25.56	65.58	23.48	**25.05**	**4.82**
Group C: MyST_train + CMU_train Finetuning:									
14	Medium	12.14	41.83	4.46	158.75	113.07	125.05	33.24	6.10
15	Medium.en	**12.10**	31.29	**2.27**	138.95	125.37	77.38	33.32	6.13
16	Large-V2	12.37	**23.62**	2.32	184.24	211.01	180.79	48.34	**4.81**
Group D: MyST_train + PF_br_train Finetuning:									
17	Medium	**12.22**	**2.98**	16.05	**16.52**	**51.53**	14.08	**22.80**	5.40
18	Medium.en	12.33	3.32	**15.08**	17.48	59.94	**13.95**	23.41	**4.88**
19	Large-V2	13.34	4.17	17.11	26.55	58.37	20.24	24.94	4.97
Group E: MyST_train + CMU_train + PF_br_train Finetuning:									
20	Medium	11.72	3.11	2.36	23.94	86.13	16.72	27.88	5.62
21	Medium.en	**11.71**	**3.02**	2.23	**21.65**	**68.10**	**15.87**	**26.43**	5.57
22	Large-V2	12.37	3.10	**1.86**	43.34	71.18	56.29	32.99	**4.75**
Group F: MyST_train + PF_br_train + NN_10 Finetuning:									
23	Medium	**11.73**	**3.15**	**9.33**	9.12	34.59	**5.10**	**16.02**	5.33
24	Medium.en	11.81	3.36	9.58	10.37	35.27	6.22	17.04	4.95
25	Large-V2	12.75	7.05	9.71	**8.39**	**33.48**	5.63	16.67	**5.09**
Group G: MyST_train + PF_br_train + NN_20 Finetuning:									
26	Medium	11.96	3.12	**8.92**	7.74	36.21	4.16	14.40	5.39
27	Medium.en	12.30	3.28	9.53	8.94	34.78	4.42	14.87	5.01
28	Large-V2	**11.60**	**3.09**	9.22	**7.24**	**31.46**	**3.98**	**13.83**	**4.47**
Group H: MyST_train + CMU_train + PF_br_train + NN_10 Finetuning:									
29	Medium	12.75	**3.11**	**1.98**	**8.99**	36.67	**5.14**	16.09	6.09
30	Medium.en	12.35	3.42	2.06	9.04	35.92	5.84	17.55	5.28
31	Large-V2	**11.73**	3.13	2.56	9.67	**35.05**	5.51	**15.83**	**4.69**
Group I: MyST_train + CMU_train + PF_br_train + NN_20 Finetuning:									
31	Medium	12.55	3.09	1.96	**7.66**	34.77	**4.11**	**14.31**	6.06
33	Medium.en	11.88	3.28	1.98	8.16	34.99	4.65	15.87	5.15
34	Large-V2	**11.62**	**2.84**	**1.75**	8.36	**34.26**	4.40	14.52	**4.53**

It can be observed from Table 21 that larger whisper models generally performed better in recognizing speech. English-only models outperformed multilingual models, suggesting the benefits of language-specific training. Finetuning the models with MyST_train improved ASR results across various test datasets, yet this was not the case with CMU_test. The integration of CMU_train in the finetuning process unexpectedly raised WERs on child speech datasets, suggesting a closer acoustic similarity to adult speech. A pivot to using the PF_br_train dataset for finetuning improved ASR performance on non-native child speech datasets. Moreover, the inclusion of a small non-native dataset (NN_10) markedly enhanced the ASR performance on non-native datasets, and expanding this dataset to NN_20 led to further improvements. The addition of CMU_train had a neutral effect on non-native datasets in later experiments, indicating that a diverse training mix could mitigate potential negative impacts of specific datasets. A comprehensive analysis and discussion of these results are thoroughly documented in our published paper, the full text of which is included in Appendix E.

5.4.1 Comparison with Previous SOTA Results

Table 22 showcases our results on various test sets alongside those from previous studies, highlighting significant improvements. However, due to differences in data cleaning methodologies used by prior researchers and the lack of a standardized process, a direct comparison is challenging. We aim to demonstrate the effectiveness of our methodology rather than directly benchmarking against prior work, acknowledging the variations in data preprocessing practices. We achieved relative WER improvements of 29.7% on the MyST_test, 41.5% on the PF_br_test, 89.1% on the CMU_test, and 85.1% on the PF_sw_test. Additionally, Table 22 includes results from other studies (marked in blue) that focused on whisper finetuning with different volumes of the MyST dataset, offering a perspective on the impact of dataset volume on finetuning outcomes.

Table 22: Comparison Between Our Results and Previously Reported Results on Non-Native Child Speech Datasets

Test Data	Approach [training data]	WER	Relative WER improvement
MyST_test	**Ours [MyST_train: 55 hrs]** DRAFT-SSL [240 hrs] [148] Whisper-Medium [55 hrs] [154] Whisper-Medium [125 hrs] [154]	**11.62** 16.53 14.40 8.61	29.7% over non-whisper models
PF_br_test	**Ours [PF_br_train: 10 hrs]** Filter-based discriminative autoencoder [8.4 hrs] [155] wav2vec2-SSL [7.4 hrs] [152]	**2.84** 18.77 4.86	41.5%
CMU_test	**Ours [CMU_train: 7 hrs]** TDNN-F [54.90 hrs] [152], HMM-DNN [6.34 hrs] [138] TDNN-HMM [6.34 hrs] [156] Encoder-Decoder VC [7.28 hrs] [157]	**1.75** 16.00 19.67 19.80 21.51	89.1%
PF_sw_test	**Ours [PF_sw_train: 0.24 hrs]** HMM-DNN [4 hrs] [138]	**7.66** 51.58	85.1%
PF_ge_test	**Ours [PF_ge_train: 0.68 hrs]**	**34.26**	NA
PF_it_test	**Ours [PF_it_train: 0.7 hrs]**	**4.11**	NA
SO_test	**Ours [SO_train: 0.48 hrs]**	**14.31**	NA

Note: These results are provided to show comparisons on the same datasets. Dataset distributions used for training/testing will vary in these papers. Our results use an 8:2 split for testing: training uses only a small

percentage of data for training as compared to other papers mentioned. We did not find previously reported results on PF_ge_test, PF_it_test, and SO_test datasets distribution.

The complete set of detailed experiments is available in our published paper titled ‘Exploring Native and Non-Native English Child Speech Recognition with Whisper’, which is currently under review in IEEE Access. However, a copy of the submitted paper is made available in Appendix E.

5.5 Conclusion and Final Remarks

Our comprehensive study across various ASR models—wav2vec2, Whisper, and Conformer—on child speech recognition has yielded insightful findings, guiding future ASR development. The wav2vec2 model demonstrated exceptional adaptability, significantly enhancing ASR performance with as little as 10 hours of child speech data. This improvement was even more pronounced when the model was finetuned with a mix of datasets, underscoring the importance of cross-domain data for optimal results. The BASE configuration of wav2vec2, pretrained with fewer data, proved effective in low-data scenarios. Whereas LARGE configuration, despite its extensive data requirement, offered only marginal improvements, thus not justifying the increased computational resources. Significant domain variations were also observed between the MyST, CMU_Kids, and PFSTAR datasets.

In contrast, Whisper models, while benefitting from dataset similarity during finetuning, struggled with certain child speech properties, as seen in increased WER on the CMU Kids dataset. However, when exploring non-native English datasets, Whisper models displayed improved generalization, particularly with diverse linguistic features, enhancing ASR performance on non-native child speech without showing any catastrophic forgetting on adult speech accuracy, This indicates a robustness in Whisper models, capable of adapting to varied accents and linguistic nuances.

The Conformer-transducer models, despite their potential in low-resource settings, did not surpass the finetuned wav2vec2 and Whisper models in child speech recognition. Larger Conformer models faced challenges with noisier data, although diverse finetuning datasets slightly reduced these issues. Notably, at smaller scales, Conformer models showed promise in non-finetuned evaluations, suggesting a potential niche for their application.

Overall, our findings show that the wav2vec2 model is the most effective for child speech ASR, particularly when finetuned with a combination of child speech datasets. This highlights the critical importance of dataset diversity and model scalability in enhancing ASR systems for child speech. The insights gained from this study pave the way for future research, emphasizing the need for efficient, adaptable models that can handle the variability and complexity of child speech for improved recognition accuracy.

The experimental framework described in this Chapter for finetuning is currently being used by Andrei Barcovschi (mentioned in section 5.3) for carrying out further ASR experiments for his research. The trained ASR models will also be used by Xperi to annotate the child speech data collected under the DAVID project (will be discussed in detail in Chapter 8).

Chapter 6

Contribution to Data Augmentation and Synthetic Speech Dataset Generation

In this chapter, we will talk about the Synthetic Dataset generated in this research using Augmentation and TTS methods. The aim was to provide a controllable way to generate more child speech datasets which can be used to further enhance the area of child speech research [43], [130], [156].

6.1 Using Fastpitch TTS for Child Speech Synthesis

The methodology of the synthetic child speech dataset created using Fastpitch was outlined in Chapter 4. Using this methodology, we have generated two distinct synthetic child speech datasets and a comprehensive description of the dataset demographics is presented in Table 23. These datasets are openly accessible via our GitHub repository[1]. We have provided datasets in sampling rates of 24khz and 16khz for research and analysis purposes.

Table 23: Synthetic Dataset Demographics

Dataset	*Speakers*	*Hours*	*Utterances*	*data/speaker*
CS_HS	40	29.02	28,800	43.53 minutes
CS_LJ	2	47.61	26,200	23.8 hours

1. **CS_HS Dataset:** This dataset, denoted as CS_HS was created using Harvard Sentences as the textual reference. It comprises 40 speakers selected from the LibriTTS dataset based on the highest data contributions in terms of hours.

2. **CS_LJ Dataset:** The CS_LJ dataset, aimed at generating synthetic child speech, employed LJ Speech transcripts as the textual reference. From the LibriTTS dataset, we identified one male and one female speaker with the most extensive training data for this purpose.

6.2 Adult Speech to Child Speech Augmentation

This research was undertaken by a fellow postdoc in the DAVID project, Mariam Yiwere, who proposed a novel approach to transform existing adult speech datasets into synthetic child-like speech, addressing the limited availability of children's speech datasets [35]. In order to establish an augmentation pipeline, it is essential to employ specialized tools for pitch modification and control of speech sample duration. We have opted to utilize the

[1] https://github.com/C3Imaging/child_tts_fastpitch

Combinatorial Expressive Speech Engine (CLEESE) [158] for implementing these augmentations. It is a Phase Vocoder-Based Toolbox for manipulating adult speech files by tuning their pitch and duration, making them sound more childlike. CLEESE was chosen due to its unique blend of user-friendliness and adaptability, permitting precise transformations to be applied to specific segments of input speech samples as needed.

The augmentation pipeline first filters adult speakers by comparing their voice embeddings to those of children, selecting speakers based on similarity scores. Selected speakers undergo pitch-shifting, adjusting the pitch of entire utterances using a single breakpoint function for each. Next, time-stretching is applied to these modified utterances, with precise timing derived from a forced alignment system that aligns speech with text, identifying pauses for differential stretching. This process involves creating breakpoint functions that dictate the extent of stretching for words and pauses. This streamlined process is depicted in Figure 21, showcasing the augmentation pipeline.

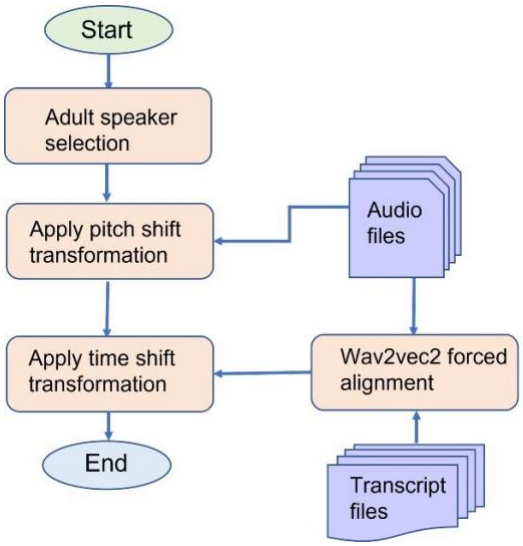


Figure 21: Flow diagram for the adult-to-child speech augmentation process.

6.2.1 AUGMENTED CHILD SPEECH DATASETS

This research resulted in creating 2 sets of synthetic child speech - Augmented_17h and Augmented_311h, using pitch-shifted and time-stretched adult speech utterances from Librispeech. The dataset details can be seen in Table 24.

Table 24: Dataset Demographics for Augmented Child Speech Datasets Using CLEESE

Dataset	Hours	Comments
Augmented_17h	17	Comprised augmented utterances from 16 female speakers in the Librispeech train-clean-100 dataset
Augmented_311h	311	Contained augmented utterances of all female speakers in the Librispeech train-clean-360, train-clean-100, dev, and test sets

The dataset created using the augmentation is not available for public distribution due to licensing restrictions with the initial datasets used in this study. However, private access may be granted upon request to interested parties who have successfully obtained their licenses for the datasets utilized, by providing proof of the obtained licenses.

6.2.2 CONTRIBUTIONS

The augmentation techniques were implemented by Mariam as a part of her research contributions to the project. My role, on the other hand, was centred around crafting the evaluation methodology for both subjective and objective assessments, as well as creating and structuring the synthetic datasets which are elucidated as follows:

Subjective Evaluations of Synthetic Speech: The study included subjective evaluations through Mean Opinion Score (MOS) tests, assessing the convincingness and intelligibility of the synthetic child's speech. The subjective study was based on the previously proposed subjective MOS study in Chapter 4, however, it was modified as required for this research. The goal of this MOS study was to determine: i) the ideal pitch-shift value for each speaker, ii) how realistic or convincing the augmented utterances sound and iii) Evaluate if the augmented utterances retain intelligibility or become unclear. Therefore, the MOS study was conducted in two phases with 60 evaluators to determine these factors. We present the results of the MOS study in Table 25 for Convincingness and Intelligibility.

Table 25: Mean and Standard Deviation (Std) of Convincingness and Intelligibility MOS Scores (C-MOS and I-MOS) From the MOS Study

Speakers	Count	C-MOS (STD)	I-MOS (STD)
Female	16	3.37 (0.37)	4.32 (0.20)
Male	4	1.76 (0.37)	3.87 (0.39)
All	20	3.05 (0.75)	4.23 (0.30)

Empirical Testing for Optimal Pitch-Shift Factors: The study conducted empirical tests to determine the most effective pitch-shift factors for male and female speakers, providing insights into gender-specific differences in speech augmentation.

Speaker Embedding Visualization: Uniform Manifold Approximation and Projection for Dimensionality Reduction (UMAP) [159] was used to plot speaker embeddings of both adults and children, aiming to identify adult speakers closest to the children's embeddings. In Figure 22, these embeddings are visually represented, with male speakers marked by black crosses, female speakers by blue triangles, and child speakers by red circles. Notably, child speakers' embeddings clustered in a distinct area of the space. However, visually determining the closest adult speakers to children proved difficult, leading to the adoption of a cosine similarity-based approach for more precise identification in the augmentation experiments.

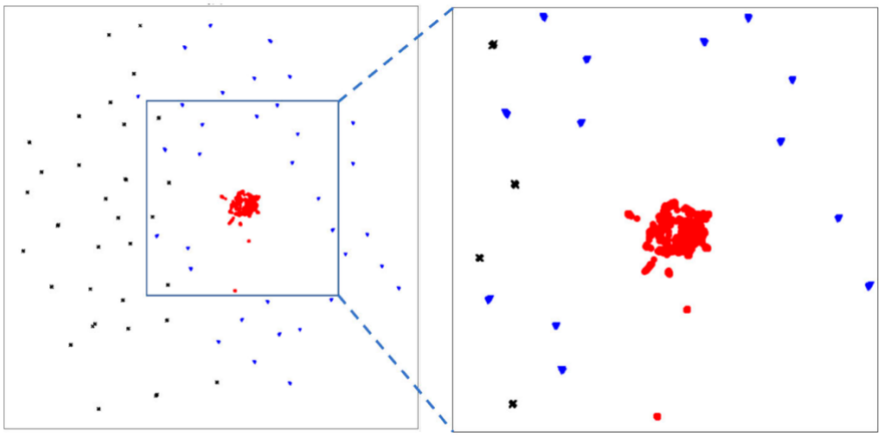


Figure 22: T-SNE Projection of 65 adult speaker embeddings from Librispeech: 31 male (black), 34 females (blue) and 31 child speaker embeddings from CMU_Kids.

Objective Evaluation for Speaker Similarity: Cosine similarity effectively measured the closeness of adult speaker embeddings to the average child speaker embedding. Consequently, for a comprehensive assessment of the augmented speech, we recalculated these cosine similarities by comparing speaker embeddings pre-augmentation and post-augmentation. Figure 23 illustrates the cosine similarities for different male and female speakers before and after the augmentation. This recalculation revealed a general increase in similarity values for all speakers. It's noteworthy that the cosine similarities between individual child speakers' embeddings and the mean child embedding ranged from 0.9 to 0.973, with one exception at 0.837.

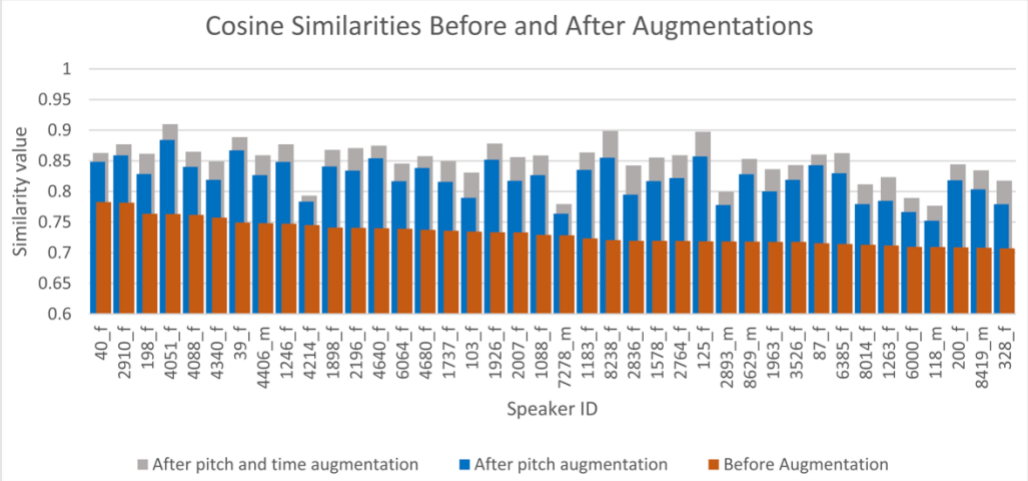


Figure 23: Cosine similarities between adult and child speaker embeddings before and after pitch shifting and time stretching augmentations.

Objective Validation of Synthetic Speech Through wav2vec2 ASR: The wav2vec2 ASR framework from Chapter 5 was used to provide the validation of the synthetic child speech dataset. The augmented speech datasets, **Augmented_17h** and **Augmented_311h** were used to finetune the wav2vec2 ASR model. The complete list of experiments can be seen in Table 26. The Librispeech dataset with BASE configuration was used for all the pretraining. Finetuning involved various datasets: Original_12h and Original_220h are the non-augmented versions of Augmented_17h and Augmented_311h, while MyST_55h was introduced in previous chapters 3 and 4. The results showed modest improvements in ASR performance when using augmented speech compared to only using adult speech, validating the effectiveness of the augmented child speech data.

Table 26: WER of wav2vec2 Models Finetuned with Original and Synthetic Speech

Model ID	Group	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMU_KIDS	WER dev_clean
1	A	Original_12h	19.95	25.10	18.95	5.78
2		Augmented_17h	20.11	**20.48**	19.14	6.58
3		Original_12h + Augmented_17h	**18.17**	**18.11**	**16.07**	5.49
4	B	MyST_55h	8.13	17.67	16.47	7.72
5		MyST_55h+ Original_12h	8.10	16.76	15.45	5.62
6		MyST_55h+ Augmented_17h	**7.98**	**14.02**	**15.02**	**4.87**
8	C	Original_220h	15.09	16.59	14.41	4.39

9		Augmented_311h	17.42	**15.86**	15.09	4.83

The complete working methodology along with the detailed experimental results is presented and published in IEEE Access titled “Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale”. A copy of this published paper is attached and presented in Appendix G of this report.

6.3 Conclusion and Final Remarks

These methodologies highlight different approaches to synthesizing child speech: one by transforming adult speech and the other by synthesizing child speech data from TTS models. The augmented synthetic datasets (Augmented_17h and Augmented_311h) were used to finetune ASR models. These models showed notable improvements in recognizing real child speech compared to models trained only on adult speech. The synthetic datasets created using Fastpitch were subjected to objective assessments for naturalness, intelligibility, and speaker similarity, which demonstrated a significant correlation between real and synthetic child voices. The synthetic child speech datasets serve as a valuable resource for further research in child speech synthesis and can potentially be used to train and improve TTS and ASR systems, especially in scenarios where real child speech data is scarce. These datasets have been made publicly available for research purposes. By making these datasets accessible, we aim to facilitate and encourage further exploration and development within this area of research. The availability of these datasets is expected to contribute significantly to the collective knowledge and progress in this domain, fostering collaboration and shared learning among researchers worldwide.

Chapter 7

Additional Contributions

7.1 Contribution to Speech Technology and Human-Computer Dialogue Conference (SpeD 23) Special Session

The special session "Research Advances in Child Speech Technologies," was organized by my supervisor Peter Corcoran and myself at the University Politehnics of Bucharest, Romania from October 25th, 2023, to 27th October 2023 for the SpeD 23 [160] conference. It focused on several key areas in the domain of child speech technology. The primary objectives of the session included exploring applications of speech synthesis and data augmentation to enhance ASR and TTS technologies for children, addressing the limitations of traditional supervised learning in low-resource child speech languages, and investigating the potential of self-supervised and unsupervised learning methods. Another significant topic was the use of transfer learning and finetuning methodologies in TTS to create synthetic child speech datasets. Additionally, the session delved into the challenges and advancements in audio-visual facial animation for children. The overview highlighted the rapid progress in deep learning for speech technology, but also the stagnation in child speech due to data scarcity and the unique differences between adult and child speech characteristics. The session aimed to unite researchers to discuss these critical technologies, their current limitations, and potential solutions for advancing speech technologies in the context of child speech. The topics of interest covered a broad range, including speech augmentation, recognition of child speech as a low-resource area, synthetic speech tools, methodologies for child speech technology enhancement, and audio-visual facial animation for children.

I had the privilege of overseeing a series of engaging and insightful presentations, each contributing significantly to the field of audio technology, particularly in the context of child-focused applications. My responsibilities involved planning and coordination to ensure the session's success, conceptualizing the session's theme to facilitating discussions and reviewing the papers for the conference. Overall, overseeing the special session and preparing the associated paper was a rewarding experience that complemented my research work and contributed to my professional development in the academic field. The session's success was a testament to the collaborative efforts of the speakers and the relevance of their research in today's rapidly evolving technological landscape.

7.2 Contribution to the DAVID Smart-Toy Platform Project

As an integral part of the DAVID project, my role was dedicated to facilitating the integration of TTS and ASR technologies onto the Ergo platform in collaboration with Xperi's engineering teams. This endeavour involved close cooperation with cross-functional experts, including engineers, linguists, and product managers, both at Xperi

Ireland and Xperi USA. Figure 24 represents a high-level system hardware architecture for the DAVID project, also showcasing the use of the ERGO chip.

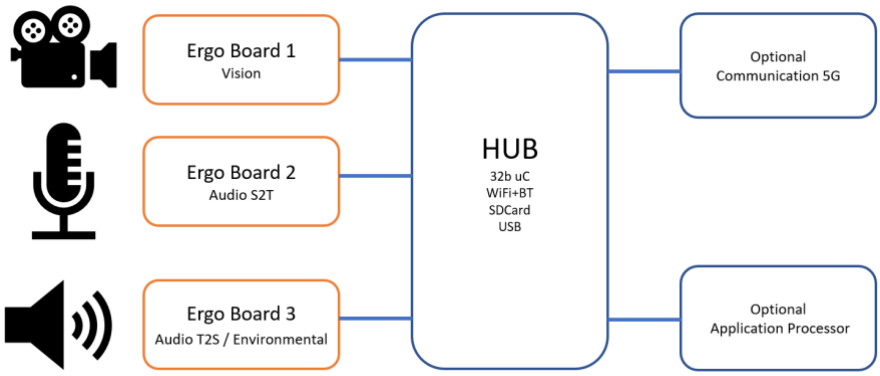


Figure 24: System hardware architecture for DAVID.

My responsibilities encompassed several critical aspects of this research. I diligently documented research findings, methodologies, and experimental results. This documentation served as a valuable resource for tracking progress, sharing insights, and ensuring the transparency of our research efforts. I also undertook the design and development of algorithms tailored specifically for speech recognition and speech synthesis on edge devices. These algorithms were engineered to meet the unique constraints and requirements of the Ergo platform, optimizing performance and efficiency. Finally, collaborative engagement with Xperi's engineering teams was a fundamental aspect of my role. Together, we worked on finetuning and optimizing AI models to ensure their seamless deployment in a production environment. This optimization process was crucial for achieving the desired real-world performance and reliability of the TTS and STT technologies. In summary, my involvement in the DAVID project entailed a multifaceted collaboration, from algorithmic development and optimization to cross-functional teamwork, ultimately contributing to the successful integration of cutting-edge speech technologies into the Ergo platform for the development of the DAVID smart toy. My efforts were also focused on the preparation and editing of a paper, ensuring the paper not only aligned with the conference's standards but also effectively conveyed the novel insights and findings relevant to the session's focus.

The DAVID platform represents a significant step forward in the practical application of Edge-AI in consumer products, particularly in the domain of smart toys. By integrating advanced neural models directly with sensory data sources and emphasizing data privacy and security, the platform sets a new standard for smart-toy development. Its flexible architecture allows for diverse applications, ranging from interactive storytelling to complex user interaction scenarios, all while adhering to strict data privacy standards. The innovative use of low-power, high-efficiency AI models demonstrates the potential for broader applications of Edge-AI in consumer electronics.

Figure 25: DAVID smart toy demo for proof-of-work depiction.

Figure 25 presents a proof-of-work demo of the DAVID smart toy by moving its eyes and providing interaction. The published paper in Sped23 titled "Data Center Audio/Video Intelligence on Device (DAVID) - An Edge-AI Platform for Smart-Toys," where I am a co-author provides an in-depth exploration of a pioneering Edge-AI platform specifically designed for smart toys. The published paper is made available in Appendix H.

7.3 Contribution to Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing

This work was undertaken by a fellow PhD student, Dan Bigioi. It presents a detailed study on creating a novel pipeline for animating facial landmarks in response to speech, which is particularly relevant for the task of automated dubbing [37]. The primary objective was to create a novel neural pipeline that could generate 3D animated facial landmarks synchronized to a target speech signal. This is particularly aimed at automating the dubbing process. A key goal was to ensure that the generated landmarks were aware of the head pose and identity characteristics from a given video, maintaining the quality of the original performance. The research aimed at devising efficient data processing methods for landmark extraction and audio feature preparation. Figure 26 shows a high-level overview of the speech-driven facial animation pipeline for automated dubbing.

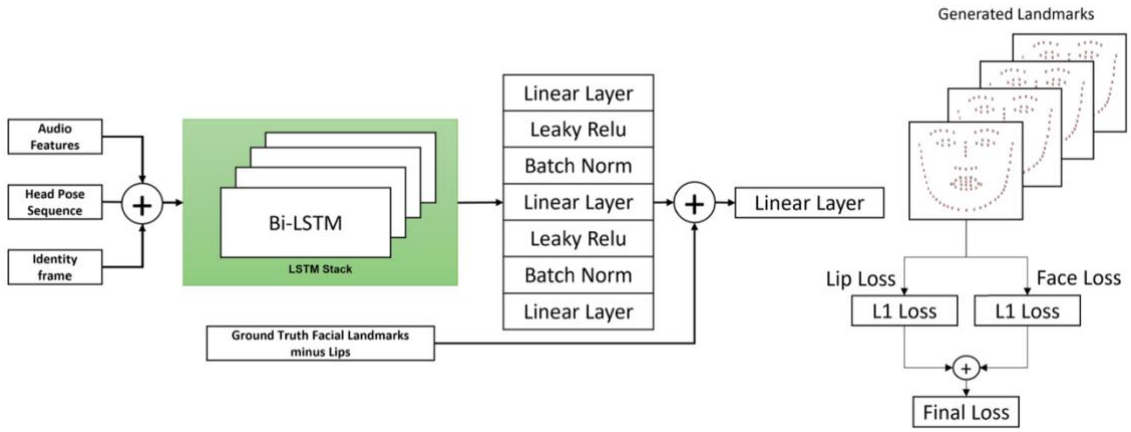


Figure 26: High-level architecture of pose-aware speech driven facial landmark animation pipeline [37].

I worked on developing and processing the speech-driven features. This entailed extracting Mel Coefficients from speech sequences, which are fundamental in driving the animation of facial landmarks. This process involves audio processing techniques and a deep understanding of how speech features correlate with facial movements, especially in the context of automated dubbing. I was also involved in preparing and conducting the MOS study. This includes designing the study's framework, selecting appropriate video samples,

and determining the evaluation criteria. A well-structured MOS study is essential for objectively assessing the quality of the generated landmarks and the overall effectiveness of the model. Finally, I helped in setting up and managing multi-GPU training environments significantly accelerating the training process, allowing for more efficient experimentation and iteration. Since multi-GPU training often involves challenges such as ensuring efficient data parallelism and balancing loads across GPUs, I also contributed to troubleshooting these issues and optimizing the training process. In summary, my contributions as a supporting researcher included the development of speech-audio features, preparation and advising on design elements of the MOS study, setting up experiments in multi-GPU server and troubleshooting and optimizations of these experiments. This study's contribution lies in its novel approach to integrating speech and audio features with facial landmark generation, enhancing the realism and accuracy of automated dubbing and facial animation. It shows a method for creating 3D pose and identity-aware talking head landmarks from a source video and driving speech signal. The use of LSTM-based networks and Mel Coefficients, along with effective training methodologies, stands out as a significant advancement in the field. The introduction of a novel LSTM-based model and Procrustes Lip Augmentation technique for data processing significantly contributes to the field, providing a solid foundation for future research.

The "Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing," provides a critical contributions to the development of advanced synthetic speech-driven facial animation technologies. Furthermore, it addresses GDPR privacy and ethical considerations by facilitating the generation of synthetic training data, thus reducing the dependency on extensive real data collection from children. The MOS study used in this pipeline is a more comprehensive version of the one mentioned in section 4.1.3, modified to align with the requirements of the facial animation subjective evaluation. This synthetic data generation capability is vital for developing and testing machine learning models in child speech research. Additionally, the integration of this facial animation technology with synthetic voice generation techniques significantly enhances human-computer interaction by enabling more natural and engaging interactions with digital avatars. This advancement is crucial for applications such as educational tools, smart toys and therapeutic platforms designed for children. The application of these technologies to create realistic and believable synthetic speaking children avatars will be discussed in more detail in section 7.4.

The published paper in IEEE Access titled " Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing" provides more insights into this pipeline. The published paper is also made available in Appendix I of this thesis.

7.4 Contribution to Synthetic Speaking Children – Why We Need Them and How to Make Them

This study represents the final contribution to the DAVID project. It seamlessly integrates three distinct research works: **'ChildGAN: Large Scale Synthetic Child Facial Data Using Domain Adaptation in StyleGAN'** [161] by Mohammad Ali Farooq and Wang Yao, **'Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing'** [37] by Dan Bigioi, as discussed in Section 7.2, and the **Text-to-Speech (TTS) research** [29], [30] conducted in Chapter 4. This study represents a key contribution to our supervisor's experiment, merging three distinct PhD projects to address the restrictions on public dissemination of real children's facial data due to Data Protection

Obligations (DPO). Collectively, this research represents state-of-the-art work in the creation of synthetic-speaking children, with a specific focus on applications within the audio-visual multimodality domain. This achievement gains particular significance when viewed in the context of an Edge-AI smart-toy platform by enabling the development of synthetic datasets.

The methodology outlined in the research represents a comprehensive approach to creating synthetic-speaking children, combining techniques in face generation, speech synthesis, and facial animation. Figure 27 represents the pipeline designed for generating 3D synthetic-speaking children. Each step contributes uniquely to the final goal, showcasing innovative uses of AI and generative models in synthesizing realistic children's faces and voices. The initial step involved transforming synthetic adult faces generated by StyleGAN [162] into child-like faces. This process involves morphological alterations and age regression techniques, for creating a base dataset (Seed Data) representative of children. Following the initial transformation, the StyleGAN generator was finetuned specifically to generate synthetic child faces (see Figure 28). This finetuning involved retraining the network with the 'childified' seed data to better capture the unique attributes of children's faces. The research then focused on synthesizing child speech using text-to-speech (TTS) models like FastPitch and Tacotron2. Techniques like Cleese-based pitch augmentation were employed to modify adult speech data to resemble that of children, achieving realism in voice timbre and prosody. Finally, the generated child faces and synthesized speech were brought together using a Pose-Aware facial animation pipeline. This step involved animating the synthetic faces to match the TTS-generated speech, ensuring lip synchronization and expressive facial movements. This demonstrates a sophisticated understanding of both audio-visual synchronization and facial animation technology, culminating in the creation of lifelike synthetic speaking children.

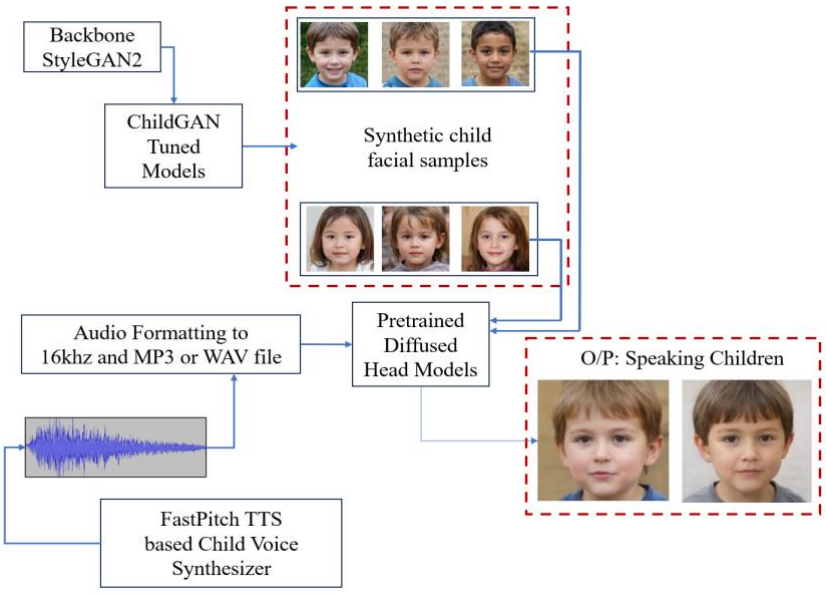


Figure 27: Block diagram representing the pipeline adapted for generating 3D synthetic child-speaking clips.

The research methodology employed in this study is characterized by the integration of three distinct studies, which inherently presents replication challenges. To ensure accurate and feasible replication, detailed guidelines and all necessary materials are available on the paper's GitHub repository, thereby facilitating a smoother replication process for researchers and practitioners interested in exploring or building upon this work. The

evaluation of the synthetic data's quality included a subjective assessment by six members of our research group. They were asked to rate the visual and audio quality of the synthetic child video, and whether the overall video appeared natural and sharp. Five participants positively responded to the video's visual quality and its overall naturalness, while four agreed on the good audio quality. This resulted in an average positive response rate of 75% from the human evaluators.

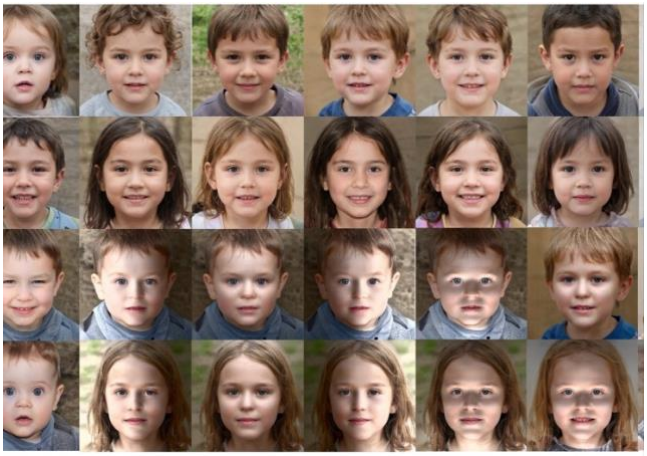


Figure 28: Distinct child facial samples of boys and girls generated from ChildGAN which were passed through this pipeline to generate synthetic-speaking children [161].

In this research, my primary contributions revolved around the development of TTS synthesis for child speech. Specifically, I worked towards integrating cutting-edge technologies like Fastpitch and Tacotron 2 TTS, leveraging advanced voice augmentation techniques to create genuine children's voices that capture their unique characteristics. Additionally, I was also involved with combining the generative TTS models with the 3D landmark-based talking heads pipeline. Furthermore, I helped with the paper-writing process for this research. This was a really enjoyable piece of research. I am pleased that I had the opportunity to integrate my research contributions with those of others, resulting in the creation of innovative knowledge within our field of study. This research can revolutionize the way synthetic data is used in various domains, offering more realistic, diverse, and ethically sourced data for a wide range of applications. The complete work is presented in the published paper ‘Synthetic Speaking Children – Why We Need Them and How to Make Them’, a copy of which is made available in Appendix J.

A notable aspect of this study is the remarkably realistic appearance of the synthetic children it produces. This realism, resulting from the integration of diverse research methods, not only broadens the study's applicability but also highlights advancements in synthetic human representation technologies. This successful creation of lifelike synthetic children paves the way for ethical research in areas where using real child data is limited.

Chapter 8

Conclusion and Future Work

In this thesis, we collectively make significant contributions to the advancement of AI-based ASR and TTS systems for child speech understanding. We comprehensively cover the evolution of speech technology, emphasizing the synergy between TTS and ASR in tackling the unique challenges of child speech. Furthermore, they underscore the critical need for ethically sound child speech datasets and showcase substantial progress in developing TTS and ASR technologies tailored for children. The exploration of synthetic datasets also demonstrates their potential utility in situations with limited real-world child speech data. we address the problem of the low-resource nature of child speech and provide solutions to encounter and improve this area of research.

8.1 Summary of the Contributions Presented in This Thesis

Reflecting on my involvement in this comprehensive research, I can affirmatively say that it has been an immensely positive and enriching experience. The work presented in each chapter has not only contributed significantly to the field of speech technology but has also provided me with invaluable skills and insights. In this section, a concise summary of the main contributions of this thesis is presented for each of the chapters:

Chapter 2 outlines the progression, mechanisms and selection of speech technology, emphasizing TTS and ASR. The chapter highlights the synergistic relationship between TTS and ASR, particularly in addressing the unique challenges of child speech. It then traces TTS from early synthesis methods to advanced neural models like Tacotron and Fastpitch, detailing their structures and functionalities. In ASR, the chapter reviews historical milestones, from Bell Laboratories' early systems to contemporary deep learning models like wav2vec2, Whisper, and Conformer. It also delves into the selection and evaluation of ASR and TTS models used in this work.

Chapter 3 of the study addresses the creation of child speech datasets, vital for enhancing TTS and ASR for children. It discusses the scarcity of such datasets and their impact on TTS and ASR research, emphasizing the challenges in accurately representing and transcribing children's unique speech characteristics. Furthermore, it describes the processes of cleaning and preprocessing these datasets to overcome limitations in data availability and variability. This chapter provides a comprehensive data-cleaning methodology, designed to establish a benchmark standard for processing and working with child speech data. This standardized approach to data cleaning is expected to facilitate better quality research and comparability across different studies focusing on child speech.

Chapter 4 details advancements in TTS technology for child speech within the DAVID project, utilizing Tacotron2 and Fastpitch models in a transfer learning framework. The

chapter outlines the development and optimization of these models using cleaned child speech datasets, focusing on generating realistic synthetic child speech. Evaluations, both subjective and objective, indicate that these models effectively produce child-like speech, though some limitations in naturalness and intelligibility persist. This research marks significant progress in creating diverse and natural child voices for TTS applications, highlighting the ongoing need for more varied child speech datasets for further improvements.

Chapter 5 focuses on enhancing ASR for child speech. The study addresses the challenges of ASR with child speech. Through extensive experiments, the chapter demonstrates that wav2vec2, when fine-tuned with even a small amount of child speech data, significantly improves ASR performance. The chapter also explores the adaptation of Whisper models for child speech and compares their effectiveness with wav2vec2 and Conformer Transducer models, providing insights into the development of more efficient ASR systems tailored for child speech recognition.

Chapter 6 addresses the creation of synthetic child speech datasets using Fastpitch TTS and adult-to-child speech augmentation techniques. Two datasets, CS_HS and CS_LJ, were generated using Fastpitch TTS, while Augmented_17h and Augmented_311h were produced by transforming adult speech into child-like speech using CLEESE. These datasets underwent objective and subjective evaluations and were used to finetune ASR models, showing performance improvements. This chapter highlights the potential of synthetic datasets in child speech research, particularly in scenarios with limited real child speech data.

In Chapter 7, the focus is on additional contributions to the DAVID project and related research. Key involvements include collaborating with Xperi on integrating TTS and STT technologies for the DAVID smart toy, contributing to a facial landmark animation pipeline for automated dubbing, and participating in a study on synthetic-speaking children. These contributions encompass algorithm development, documentation, optimization with engineering teams, and the development of advanced speech audio features for facial animation. The work on synthetic-speaking children involved integrating text-to-speech synthesis with 3D facial animations, demonstrating the potential of AI in creating realistic synthetic data for diverse applications. The outcomes of these efforts are detailed in published papers and highlight significant advancements in speech synthesis, facial animation, and smart toy technology.

Overall, this research journey has been instrumental in shaping my skills as a researcher, particularly in the domain of speech technology. It has prepared me for future challenges in the field, contributing to my growth in technical expertise, ethical research practice, and applied research skills. In terms of the impact on the field of speech technology, this work paves the way for future innovations, especially in creating more nuanced, effective, and ethically sound speech technologies for children. The knowledge and methodologies developed in this research have the potential to influence a wide range of applications, from educational tools to interactive media, enriching the landscape of speech technology research and development for children.

8.2 Discussion and Future Work

This field of child speech research is considered a relatively unexplored domain with significant potential for advancement. Below, we delve into several considerations and limitations of the current work, as well as prospective areas for future exploration. Additionally, we discuss intriguing long-term research possibilities in this field, and highlight various research paths for follow-up studies.

8.2.1 LIMITATIONS

Although this thesis has achieved notable progress in speech technologies tailored to children aged 5 to 12, it's essential to acknowledge the diverse developmental paths observed among individuals at both the younger and older extremes of the 4 to 15 year old range. Infants and toddlers, for instance, may exhibit more disfluencies, limited vocabulary, and immature articulation patterns, requiring specialized data collection methods and model architectures to effectively capture their speech characteristics. On the other hand, older adolescents may have speech patterns that are closer to adult-like, necessitating the exploration of transfer learning techniques or the incorporation of age-specific linguistic features to ensure optimal performance.

Furthermore, the environmental and cultural factors that shape child speech can vary widely, not only across geographical regions but also within different socioeconomic, educational, and linguistic contexts. The conclusions drawn in this thesis may be most applicable to the specific settings in which the research was conducted, which may not fully represent the diversity of child speech experiences globally. Expanding the scope of data collection and model evaluation to include a broader range of cultural and environmental influences could yield valuable insights into the generalizability of the developed technologies.

Additionally, longitudinal studies tracking the speech development of individual children over time could provide deeper understanding of the dynamic nature of child speech and inform the design of adaptive speech technologies that can evolve alongside the child's linguistic and cognitive growth. By acknowledging the limitations of the current work and proactively addressing the diverse needs and characteristics of children across the age spectrum and various environmental contexts, future research in this field can build upon the foundations laid by this thesis to develop truly inclusive and responsive speech technologies for the benefit of all young users.

8.2.2 FUTURE WORK

Steps Towards Improving Child Speech TTS: To enhance multi-speaker TTS models, there is a plan to integrate further refinement of vocoders, possibly through the utilization of state-of-the-art GAN-based vocoders like HiFi-GAN [93]. This will require access to a larger dataset of child speech to achieve significant improvements. Incorporating multi-accented non-native datasets into the TTS pipeline also presents a compelling avenue for generating English child speech with various accents. This strategy acknowledges the diversity inherent in language expression among children from different geographical and cultural backgrounds. By training TTS models on a variety of accents, the resultant speech synthesis could reflect the nuances and subtleties of regional pronunciations. To achieve this, a two-stage finetuning process can be adopted. Initially, a TTS model would undergo finetuning with a multispeaker child TTS dataset, which would lay the groundwork for broad accent coverage. The second stage of finetuning would focus the model on a single,

specific accent. This hierarchical approach allows the model to first acquire a wide range of phonetic and prosodic patterns before honing in on the unique characteristics of a particular accent. Preliminary tests of this method have yielded promising results, indicating the viability of this approach.

Subjective Evaluation: Having gained considerable experience in conducting Mean Opinion Score (MOS) studies, a subjective evaluation of the generated child speech datasets from Chapter 6 will be conducted to assess naturalness, intelligibility, and speaker similarity more comprehensively. The challenges posed by Mean Opinion Score (MOS) studies, especially regarding child speech, stem from the absence of a standardized benchmark. Moreover, subjective evaluations inherent to MOS studies are inherently resource-intensive, demanding significant time and effort from both researchers and participants. Our efforts to streamline MOS methodologies for child speech can provide researchers with an easier way to conduct subjective studies as it eliminates the need for extensive time and resource investments. This methodology, which we aim to establish as the benchmark for future evaluations, will be instrumental in ensuring consistency and reliability in how child speech synthesis is appraised.

Steps Towards Gathering More Child Speech Data: Enhancing this research fundamentally hinges on the expansion and diversification of child speech datasets. In the duration of the DAVID project, Xperi accumulated a substantial quantity of unannotated child speech data across various recording environments, a topic not addressed in this thesis owing to concerns related to confidentiality and ethics. The ASR models tailored for child speech recognition represent a pivotal tool for transcribing the substantial volume of unlabelled child speech data collected by Xperi (see section 3.4.1). The application of these models to Xperi's extensive, unannotated child speech datasets will streamline the process of transforming raw audio into structured, textual data. We also intend to utilize the data gathered in section 3.4 by Xperi and Bits Pilani using the data collection application for conducting additional experiments involving child speech. This will help models generalize better to Irish and Indian-accented English child speakers. It would also help improve the ASR transcription and provide a better generalization to unannotated child speech datasets gathered by Xperi.

Steps Towards Improving Child Speech ASR: There are plans to conduct further finetuning experiments, particularly focusing on smaller Conformer-transducer models that are adapted specifically for child speech data. Such finetuning will aim to leverage the unique characteristics of child speech to enhance model performance. Moreover, the research will delve into the optimization of hyperparameters, which is a critical step toward refining the ASR models' ability to learn and generalize from the data. Experimenting with different decoding strategies will also form a key part of this.

8.2.3 LONG-TERM RESEARCH PROSPECTS

Experiments with Synthetic Speech: It would also be interesting to use the synthetically generated child speech as a data augmentation technique for ASR and speaker recognition models. This approach entails to creation of realistic, synthetic child speech data, which could vastly expand the available training resources without the ethical and practical constraints of recording real child speech [44], [130]. This can further contribute to enhancing the performance of child speech technologies. Since these datasets are available publicly, we also encourage researchers to use these synthetic datasets.

Experiments to Improve Child Speech Recognition: An additional concept under

consideration was to adapt the 'Pretraining' phase of the wav2vec2 model by utilizing a 'cleaned' dataset of child speech. This approach arises from the observation that incorporating the MyST dataset into the pretraining increased the Word Error Rate (WER). It would be intriguing to examine the effects of integrating a refined set of child speech data, one that has undergone meticulous cleaning processes to enhance its quality. Such a dataset could potentially eliminate confounding noise and variability, providing the wav2vec2 model with clearer, more consistent speech patterns during pretraining. This could lead to improved model performance, as the cleaner data may help the model learn more accurate representations of child speech. It would also be interesting to improve the baseline ASR models by incorporating additional training datasets from different low-resource languages as it could significantly bolster its performance, especially in recognizing non-native child speech. This expansion entails integrating diverse linguistic data, which may introduce a broader spectrum of phonetic and acoustic patterns, thereby improving the model's ability to generalize across various speech characteristics.

Deployment on Edge Devices: The exploration of deploying ASR models on edge devices represents a vital direction for future research. This deployment strategy involves adapting the models to function within the constraints of lower-power devices, which requires efficient model architectures that maintain performance while operating with limited computational resources. Optimizing ASR models for edge deployment could significantly reduce latency, enhance privacy by processing data locally, and ensure functionality even without constant internet connectivity.

References

[1] A. Tjandra, S. Sakti, and S. Nakamura, ‘Listening while speaking: Speech chain by deep learning’, in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, Dec. 2017, pp. 301–308. doi: 10.1109/ASRU.2017.8268950.

[2] A. Tjandra, S. Sakti and S. Nakamura, "Machine Speech Chain," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 976-989, 2020, doi: 10.1109/TASLP.2020.2977776.

[3] V. Bhardwaj *et al.*, ‘Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review’, *Appl. Sci.*, vol. 12, no. 9, 2022, doi: 10.3390/app12094419.

[4] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, ‘A review of ASR technologies for children’s speech’, in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, in WOCCI ’09. New York, NY, USA: Association for Computing Machinery, Nov. 2009, pp. 1–8. doi: 10.1145/1640377.1640384.

[5] S. Lee, A. Potamianos, and S. S. Narayanan, ‘Analysis of children’s speech: duration, pitch and formants’, in *EUROSPEECH*, 1997.

[6] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, ‘Study of formant modification for children ASR’, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7429–7433.

[7] I. Chandra Yadav and G. Pradhan, ‘Pitch and noise normalized acoustic feature for children’s ASR’, *Digit. Signal Process. Rev. J.*, vol. 109, p. 102922, 2021, doi: 10.1016/j.dsp.2020.102922.

[8] S. Lee, A. Potamianos, and S. Narayanan, ‘Acoustics of children’s speech: Developmental changes of temporal and spectral parameters’, *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999, doi: 10.1121/1.426686.

[9] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, 'Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition', In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Vol. 1, pp. I-I).

[10] R. Serizel and D. Giuliani, ‘Vocal tract length normalisation approaches to DNN-based children’s and adults’ speech recognition’, in *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2014, pp. 135–140.

[11] P. G. Shivakumar and P. Georgiou, ‘Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations’, *Comput. Speech Lang.*, vol. 63, p. 101077, 2020.

[12] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, ‘A Review of Deep Learning Based Speech Synthesis’, *Appl. Sci.*, vol. 9, no. 19, Art. no. 19, Jan. 2019, doi: 10.3390/app9194050.

[13] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, ‘A Survey on Neural Speech Synthesis’. arXiv, Jul. 23, 2021. Available: http://arxiv.org/abs/2106.15561

[14] S. Dutta *et al.*, ‘Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments Robust Speaker Diarization View project End-to-end text independent speaker recognition View project Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments’, 2022, doi: 10.21437/Interspeech.2022-555.

[15] R. Fan, Y. Zhu, J. Wang, and A. Alwan, ‘Towards Better Domain Adaptation for Self-Supervised Models: A Case Study of Child ASR’, *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1242–1252, 2022, doi: 10.1109/JSTSP.2022.3200910.

[16] A. Gulati *et al.*, ‘Conformer: Convolution-augmented Transformer for Speech Recognition’. arXiv, May 16, 2020. doi: 10.48550/arXiv.2005.08100.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, ‘wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations’, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 12449–12460.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, ‘Robust speech recognition via large-scale weak supervision’, in *International Conference on Machine Learning*, PMLR, 2023, pp. 28492–28518.

[19] Y. Wang *et al.*, 'Tacotron: Towards end-to-end speech synthesis'. *arXiv preprint arXiv:1703.10135*.

[20] Y. Jia *et al.*, ‘Transfer learning from speaker verification to multispeaker text-to-speech synthesis’, in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2018, pp. 4480–4490.

[21] A. \Lańcucki, ‘Fastpitch: Parallel text-to-speech with pitch prediction’, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6588–6592.

[22] K. Kazi, S. Devi, B. Sreedhar, and P. Arulprakash, ‘A Path Towards Child-Centric Artificial Intelligence based Education’, *Int. J. Early Child. Spec. Educ.*, vol. 14, p. 2022, Jan. 2022, doi: 10.9756/INT-JECSE/V14I3.1145.

[23] W. Yang, ‘Artificial Intelligence education for young children: Why, what, and how in curriculum design and implementation’, *Comput. Educ. Artif. Intell.*, vol. 3, p. 100061, Jan. 2022, doi: 10.1016/j.caeai.2022.100061.

[24] G. Cosache, F. Salgado, R. Jain, C. Rotariu, G. Sterpu, and P. Corcoran, ‘Data Center Audio/Video Intelligence on Device (DAVID) - An Edge-AI Platform for Smart-Toys’, in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Oct. 2023, pp. 66–71. doi: 10.1109/SpeD59241.2023.10314915.

[25] J. Chen and X. Ran, ‘Deep Learning With Edge Computing: A Review’, *Proc. IEEE*, 2019, doi: 10.1109/JPROC.2019.2921977.

[26] R. Singh and S. S. Gill, ‘Edge AI: A survey’, *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 71–92, Jan. 2023, doi: 10.1016/j.iotcps.2023.02.004.

[27] E. Li, L. Zeng, Z. Zhou, and X. Chen, ‘Edge AI: On-Demand Accelerating Deep Neural Network Inference via Edge Computing’, *IEEE Trans. Wirel. Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020, doi: 10.1109/TWC.2019.2946140.

[28] J. Shen *et al.*, ‘Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* , pp. 4779-4783.

[29] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, ‘A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis’, *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.

[30] R. Jain and P. Corcoran, ‘Improved Child Text-to-Speech Synthesis through Fastpitch-based Transfer Learning’, in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Oct. 2023, pp. 54–59. doi: 10.1109/SpeD59241.2023.10314899.

[31] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran and H. Cucu, "A WAV2VEC2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition," in IEEE Access, vol. 11, pp. 46938-46948, 2023, doi: 10.1109/ACCESS.2023.3275106.

[32] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, ‘Adaptation of Whisper models to child speech recognition’. Proc. INTERSPEECH 2023, 5242-5246, doi: 10.21437/Interspeech.2023-935.

[33] A. Barcovschi, R. Jain, and P. Corcoran, ‘A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition’, in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Oct. 2023, pp. 42–47. doi: 10.1109/SpeD59241.2023.10314867.

[34] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, ‘Exploring Native and Non-Native English Child Speech Recognition with Whisper’, *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3378738.

[35] M. Y. Yiwere, A. Barcovschi, R. Jain, H. Cucu, and P. Corcoran, ‘Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale.’, in *IEEE Access*, vol. 11, pp. 109066-109081, 2023, doi: 10.1109/ACCESS.2023.3317360.

[36] M. Ali Farooq, D. Bigioi, R. Jain, W. Yao, M. Yiwere, and P. Corcoran, ‘Synthetic Speaking Children – Why We Need Them and How to Make Them’, in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Oct. 2023, pp. 36–41. doi: 10.1109/SpeD59241.2023.10314943.

[37] D. Bigioi, H. Jordan, R. Jain, R. McDonnell, and P. Corcoran, ‘Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing’, *IEEE Access*, vol. 10, pp. 133357–133369, 2022, doi: 10.1109/ACCESS.2022.3231137.

[38] M. K. Baskar, L. Burget, S. Watanabe, and R. F. Astudillo, ‘EAT: Enhanced ASR-TTS for self-supervised speech recognition’, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6753–6757.

[39] J. Xu *et al.*, ‘LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition’, in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, in KDD ’20. New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 2802–2812. doi: 10.1145/3394486.3403331.

[40] D.-R. Liu, C.-Y. Yang, S.-L. Wu, and H.-Y. Lee, ‘Improving Unsupervised Style Transfer in end-to-end Speech Synthesis with end-to-end Speech Recognition’, in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2018, pp. 640–647. doi: 10.1109/SLT.2018.8639672.

[41] P. Alderson and V. Morrow, *The ethics of research with children and young people: A practical handbook*. Sage, 2020.

[42] T. Li, S. Yang, L. Xue, and L. Xie, ‘Controllable Emotion Transfer For End-to-End Speech Synthesis’, in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1-5. IEEE, 2021.

[43] X. Zhu and L. Xue, ‘Building a controllable expressive speech synthesis system with multiple emotion strengths’, *Cogn. Syst. Res.*, vol. 59, pp. 151–159, Jan. 2020, doi: 10.1016/j.cogsys.2019.09.009.

[44] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Data augmentation for asr using tts via a discrete representation." In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 68-75. IEEE, 2021.

[45] M. Escobar-Planas, E. Gomez, and C.-D. Martınez-Hinarejos, ‘From Ethical Guidelines to Practical Guidance to Develop Trustworthy Conversational Agents for Children’.

[46] R. Garg *et al.*, ‘The Last Decade of HCI Research on Children and Voice-based Conversational Agents’, in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, in CHI ’22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–19. doi: 10.1145/3491102.3502016.

[47] H. Dudley and T. H. Tarnoczy, ‘The Speaking Machine of Wolfgang von Kempelen’, *J. Acoust. Soc. Am.*, vol. 22, no. 2, pp. 151–166, Mar. 1950, doi: 10.1121/1.1906583.

[48] F. J. Owens, ‘Speech Synthesis’, in *Signal Processing of Speech*, F. J. Owens, Ed., in Macmillan New Electronics Series. , London: Macmillan Education UK, 1993, pp. 88–121. doi: 10.1007/978-1-349-22599-6_5.

[49] C. K. Leong, ‘Enhancing reading comprehension with text-to-speech (DECtalk) computer system’, *Read. Writ.*, vol. 4, no. 2, pp. 205–217, Jun. 1992, doi: 10.1007/BF01027492.

[50] N. Schnell, G. Peeters, S. Lemouton, P. Manoury, and X. Rodet, ‘Synthesizing a choir in real-time using Pitch Synchronous Overlap Add (PSOLA)’ In *ICMC*. 2000..

[51] O. Watts, J. Yamagishi, S. King, and K. Berkling, ‘HMM Adaptation and Voice Conversion for the Synthesis of Child Speech: A Comparison’, 2009. Available: https://era.ed.ac.uk/handle/1842/3923

[52] A. Govender, B. Nouhou, and F. De Wet, ‘HMM Adaptation for child speech synthesis using ASR data’, in *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, IEEE, 2015, pp. 178–183.

[53] H. Zen, K. Tokuda, and A. W. Black, ‘Statistical parametric speech synthesis’, *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009, doi: 10.1016/J.SPECOM.2009.04.004.

[54] M. Bulut, S. S. Narayanan, and A. K. Syrdal, ‘Expressive speech synthesis using a concatenative synthesizer’, in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, ISCA, Sep. 2002, pp. 1265–1268. doi: 10.21437/ICSLP.2002-389.

[55] S. Kayte, M. Mundada, and J. Gujrathi, ‘Hidden Markov Model based Speech Synthesis: A Review’, *Int. J. Comput. Appl.*, vol. 130, no. 3, pp. 35–39, Nov. 2015, doi: 10.5120/ijca2015906965.

[56] A. Van Den Oord *et al.*, ‘Wavenet: A generative model for raw audio.' *arXiv preprint arXiv:1609.03499* 12 (2016).

[57] S. Ö. Arık *et al.*, ‘Deep Voice 2: Multi-Speaker Neural Text-to-Speech’ in *Advances in neural information processing systems*, *30*, 2017.

[58] S. Ö. Arık *et al.*, ‘Deep Voice: Real-time Neural Text-to-Speech’, in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Jul. 2017, pp. 195–204.

[59] W. Ping *et al.*, ‘Deep voice 3: Scaling text-to-speech with convolutional sequence learning.’*arXiv preprint arXiv:1710.07654* (2017)..

[60] Y. Ren *et al.*, ‘Fastspeech: Fast, robust and controllable text to speech’, *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[61] Y. Ren *et al.*, ‘FastSpeech 2: Fast and High-Quality End-to-End Text to Speech’. *arXiv preprint arXiv:2006.04558* (2020). Available: http://arxiv.org/abs/2006.04558

[62] M. Manjutha, J. Gracy, P. Subashini, and M. Krishnaveni, ‘Automated speech recognition system—A literature review’, *Comput. Methods Commun. Tech. Inform.*, vol. 205, pp. 740–741.

[63] L. R. Rabiner, ‘A tutorial on hidden Markov models and selected applications in speech recognition’, *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989, doi: 10.1109/5.18626.

[64] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, ‘Automatic speech recognition: a survey’, *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 9411–9457, Mar. 2021, doi: 10.1007/S11042-020-10073-7/TABLES/7.

[65] S. Alharbi *et al.*, ‘Automatic Speech Recognition: Systematic Literature Review’, *IEEE Access*, vol. 9, pp. 131858–131876, 2021, doi: 10.1109/ACCESS.2021.3112535.

[66] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, ‘Listen, attend and spell: A neural network for large vocabulary conversational speech recognition’, in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 4960–4964.

[67] A. Hannun *et al.*, ‘Deep Speech: Scaling up end-to-end speech recognition’ *arXiv preprint arXiv:1412.5567* (2014).

[68] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, 'A comparison of transformer and lstm encoder decoder models for asr', in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, 2019, pp. 8-15, doi: 10.1109/ASRU46091.2019.9004025.

[69] Q. Zhang *et al.*, 'Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss', in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7829–7833.

[70] Z. Peng *et al.*, 'Conformer: Local features coupling global representations for visual recognition', in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 367–376.

[71] D. W. Griffin A N D Jae S Lim and S. Member, 'Signal Estimation from Modified Short-Time Fourier Transform', *IEEE Transactions on acoustics, speech, and signal processing* 32, no. 2: 236-243, 1984.

[72] D. Bigioi and P. Corcoran, 'Challenges for Edge-AI Implementations of Text-To-Speech Synthesis', in *2021 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2021, pp. 1–6. doi: 10.1109/ICCE50685.2021.9427679.

[73] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, 'Neural Speech Synthesis with Transformer Network', *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, Art. no. 01, Jul. 2019, doi: 10.1609/aaai.v33i01.33016706.

[74] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, 'Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis'. *arXiv preprint arXiv:2005.05957* , Jul. 16, 2020. Available: http://arxiv.org/abs/2005.05957

[75] K. Peng, W. Ping, Z. Song, and K. Zhao, 'Non-Autoregressive Neural Text-to-Speech', in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Nov. 2020, pp. 7586–7598. Available: https://proceedings.mlr.press/v119/peng20a.html

[76] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, 'Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow', in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7209–7213. doi: 10.1109/ICASSP40776.2020.9054484.

[77] J. Kim, S. Kim, J. Kong, and S. Yoon, 'Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search', *Advances in Neural Information Processing Systems* 33 (2020): 8067-8077. Available: https://github.com/jaywalnut310/glow-tts.

[78] 'The LJ Speech Dataset'. Accessed: May 24, 2021. Available: https://keithito.com/LJ-Speech-Dataset/

[79] R. Valle, J. Li, R. Prenger, and B. Catanzaro, 'Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens.' In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6189-6193. IEEE, 2020.

[80] Z. Cai, C. Zhang, and M. Li, 'From Speaker Verification to Multispeaker Speech Synthesis, Deep Transfer with Feedback Constraint', presented at the Proc. Interspeech 2020, 2020, pp. 3974–3978. doi: 10.21437/Interspeech.2020-1032.

[81] A. Kulkarni, V. Colotte, and D. Jouvet, 'Improving Latent Representation For End To End Multispeaker Expressive Text To Speech System'. Available: https://hal.archives-ouvertes.fr/hal-02978485

[82] J. Kim, J. Kong, and J. Son, 'Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech', in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 5530–5540. Available: https://proceedings.mlr.press/v139/kim21f.html

[83] F. Lux, J. Koch, and N. T. Vu, 'Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech', in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 962–969. doi: 10.1109/SLT54892.2023.10022433.

[84] J. Vainer and O. Dušek, 'SpeedySpeech: Efficient Neural Speech Synthesis'. Proc. Interspeech 2020, 3575-3579, doi: 10.21437/Interspeech.2020-2867 Available: https://github.com/janvainer/speedyspeech

[85] D. Paul, Y. Pantazis, and Y. Stylianou, 'Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions', Proc. Interspeech 2020, 235-239, doi: 10.21437/Interspeech.2020-2786.

[86] W. Jang, D. Lim, and J. Yoon, 'Universal melgan: A robust neural vocoder for high-fidelity waveform generation in multiple domains', *arXiv preprint arXiv:2011.09631*, 2020.

[87] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, 'Universal Neural Vocoding with Parallel Wavenet', in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 6044–6048. doi: 10.1109/ICASSP39728.2021.9414444.

[88] J. Lorenzo-Trueba *et al.*, 'Towards achieving robust universal neural vocoding.' *arXiv preprint arXiv:1811.06292*, 2018.

[89] A. Oord *et al.*, 'Parallel WaveNet: Fast High-Fidelity Speech Synthesis', in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, Jul. 2018, pp. 3918–3926. Available: https://proceedings.mlr.press/v80/oord18a.html

[90] R. Prenger, R. Valle, and B. Catanzaro, 'Waveglow: A flow-based generative network for speech synthesis.' in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617-3621. IEEE, 2019.

[91] K. Kumar *et al.*, 'MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis', in *Advances in neural information processing systems*, *32*, 2019.

[92] J.-M. Valin and J. Skoglund, 'LPCNET: Improving Neural Speech Synthesis through Linear Prediction', in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5891–5895. doi: 10.1109/ICASSP.2019.8682804.

[93] J. Kong, J. Kim, and J. Bae, 'HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis'. *in Advances in neural information processing systems* 33 (2020): 17022-17033. Available: https://jik876.github.io/hifi-gan-demo/

[94] N. Indurkhya, 'Natural language processing', *Comput. Handb. Third Ed. Comput. Sci. Softw. Eng.*, pp. 40-1-40–17, 2014, doi: 10.1201/b16812.

[95] P. L. Tobing and T. Toda, 'High-fidelity and low-latency universal neural vocoder based on multiband WaveRNN with data-driven linear prediction for discrete waveform modeling', *arXiv preprint arXiv:2105.09856*, 2021.

[96] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, 'Librispeech: An ASR corpus based on public domain audio books', in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.

[97] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, 'One TTS alignment to rule them all', in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6092–6096.

[98] M. Ravanelli, T. Parcollet, and Y. Bengio, 'The Pytorch-kaldi Speech Recognition Toolkit', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019, pp. 6465–6469. doi: 10.1109/ICASSP.2019.8683713.

[99] D. Amodei *et al.*, 'Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, Baidu Research-Silicon Valley AI Lab, in *International conference on machine learning*, pp. 173-182. PMLR, 2016'.

[100] Y. Wang *et al.*, 'Espresso: A Fast End-To-End Neural Speech Recognition Toolkit', in *2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings*, 2019, pp. 136–143. doi: 10.1109/ASRU46091.2019.9003968.

[101] Y.-C. Chen *et al.*, 'Speechnet: A universal modularized model for speech processing tasks', *ArXiv Prepr. ArXiv210503070*, 2021.

[102] S. Kriman *et al.*, 'Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions', in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6124–6128.

[103] J. Li *et al.*, 'Jasper: An end-to-end convolutional neural acoustic model', *ArXiv Prepr. ArXiv190403288*, 2019.

[104] R. Gretter, M. Matassoni, D. Falavigna, K. Evanini, and C. W. Leong, 'Overview of the Interspeech TLT2020 Shared Task onASR for Non-Native Children's Speech', *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-October, pp. 245–249, 2021, doi: 10.21437/INTERSPEECH.2020-2133.

[105] T. Zenkel *et al.*, 'Comparison of Decoding Strategies for CTC Acoustic Models', 2017, *arXiv preprint arXiv:1708.04469*.

[106] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, 'Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks', in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[107] D. S. Park *et al.*, 'SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition', Proc. Interspeech 2019, 2613-2617, doi: 10.21437/Interspeech.2019-2680.

[108] X. Zheng, C. Zhang, and P. C. Woodland, 'Adapting GPT, GPT-2 and BERT language models for speech recognition', in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 162–168.

[109] I. Loshchilov and F. Hutter, 'DECOUPLED WEIGHT DECAY REGULARIZATION', *arXiv preprint arXiv:1711.05101* (2017). Available: https://github.com/loshchil/AdamW-and-SGDW

[110] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, 'A survey about databases of children's speech.', in *INTERSPEECH*, 2013, pp. 2410–2414.

[111] D. Elenius and M. Blomberg, 'Comparing speech recognition for adults and children', *Proceedings of FONETIK 2004*: 156-159, 2004.

[112] C. A. Moore and B. Maassen, 'Physiologic development of speech production', *Speech Mot. Control Norm. Disord. Speech*, pp. 191–209, 2004.

[113] S. Shahnawazuddin, N. Adiga, and H. K. Kathania, 'Effect of Prosody Modification on Children's ASR', *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1749–1753, Nov. 2017, doi: 10.1109/LSP.2017.2756347.

[114] G. Yeung, R. Fan, and A. Alwan, 'Fundamental frequency feature normalization and data augmentation for child speech recognition', in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6993–6997.

[115] S. Shahnawazuddin, R. Sinha, and G. Pradhan, 'Pitch-Normalized Acoustic Features for Robust Children's Speech Recognition', *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1128–1132, Aug. 2017, doi: 10.1109/LSP.2017.2705085.

[116] J. N. Bohannon III and A. L. Marquis, 'Children's control of adult speech', *Child Dev.*, pp. 1002–1008, 1977.

[117] B. M. DePaulo and J. D. Bonvillian, 'The effect on language development of the special characteristics of speech addressed to children', *J. Psycholinguist. Res.*, vol. 7, pp. 189–211, 1978.

[118] A. R. Luria, 'A child's speech responses and the social environment', *Sov. Dev. Psychol. Anthol. N. Y. White Plains ME Sharpe 1977b*, pp. 32–64, 2017.

[119] N. Sadagopan and A. Smith, ‘Developmental changes in the effects of utterance length and complexity on speech movement variability’, 2008.

[120] G. Lindsay, ‘The collection and analysis of data on children with speech, language and communication needs: The challenge to education and health services’, *Child Lang. Teach. Ther.*, vol. 27, no. 2, pp. 135–150, 2011.

[121] A. Batliner, S. Hantke, and B. Schuller, ‘Ethics and Good Practice in Computational Paralinguistics’, *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1236–1253, Jul. 2022, doi: 10.1109/TAFFC.2020.3021015.

[122] H. Zen *et al.*, ‘LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech’, *arXiv preprint arXiv:1904.02882*, Apr. 05, 2019. Available: http://arxiv.org/abs/1904.02882

[123] C. Veaux, J. Yamagishi, and K. MacDonald, ‘CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit’, 2017.

[124] A. Nagrani, J. Son Chung, and A. Zisserman, ‘VoxCeleb: a large-scale speaker identification dataset’, *arXiv preprint arXiv:1706.08612* (2017).

[125] J. Kahn *et al.*, ‘Libri-light: A benchmark for asr with limited or no supervision’, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7669–7673.

[126] W. Ward, ‘My Science Tutor and the MyST Corpus’, *Boulder Learning Inc*, 2019.

[127] A. Batliner *et al.*, ‘The PF STAR Children’s Speech Corpus’. Proc. Interspeech 2005, 2761-2764, doi: 10.21437/Interspeech.2005-705

[128] M. Eskenazi, J. Mostow, and D. Graff, ‘The CMU kids speech corpus’, *Corpus Child. Read Speech Digit. Transcribed Two CD-ROMs Assist. Multicom Res. David Graff Publ. Linguist. Data Consort. Univ. Pa.*, 1997.

[129] J. Zhang *et al.*, ‘speechocean762: An Open-Source Non-native English Speech Corpus For Pronunciation Assessment’, *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 6, pp. 4386–4390, Apr. 2021, doi: 10.48550/arxiv.2104.01378.

[130] M.-J. Hwang, R. Yamamoto, E. Song, and J.-M. Kim, ‘TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis’, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6598–6602.

[131] L. Wan Quan Wang Alan Papir Ignacio Lopez Moreno, ‘GENERALIZED END-TO-END LOSS FOR SPEAKER VERIFICATION’. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879-4883. IEEE, 2018.

[132] M. Viswanathan and M. Viswanathan, ‘Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale’, *Comput. Speech Lang.*, vol. 19, no. 1, pp. 55–83, Jan. 2005, doi: 10.1016/J.CSL.2003.12.001.

[133] C.-C. Lo *et al.*, ‘MOSNet: Deep Learning based Objective Assessment for Voice Conversion’, in *Interspeech 2019*, Sep. 2019, pp. 1541–1545. doi: 10.21437/Interspeech.2019-2003.

[134] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, ‘CROWDMOS: AN APPROACH FOR CROWDSOURCING MEAN OPINION SCORE STUDIES’, In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2416-2419. IEEE, 2011.

[135] M. C. Cieslak, A. M. Castelfranco, V. Roncalli, P. H. Lenz, and D. K. Hartline, ‘t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis’, *Mar. Genomics*, vol. 51, p. 100723, 2020.

[136] E. Cooper *et al.*, ‘Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings.’ In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6184-6188. IEEE, 2020.

[137] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, 'Data Augmentation Using CycleGAN for End-to-End Children ASR', in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 511–515. doi: 10.23919/EUSIPCO54536.2021.9616228.

[138] T. Rolland, A. Abad, C. Cucchiarini, and H. Strik, 'Multilingual Transfer Learning for Children Automatic Speech Recognition', in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 7314–7320. Available: https://aclanthology.org/2022.lrec-1.795

[139] J. Thienpondt and K. Demuynck, 'Transfer Learning for Robust Low-Resource Children's Speech ASR with Transformers and Source-Filter Warping.' Proc. Interspeech 2022, 2213-2217, doi: 10.21437/Interspeech.2022-10964.

[140] O. Watts, J. Yamagishi, K. Berkling, and S. King, 'HMM-based synthesis of child speech', *Proc 1st Workshop Child Comput. Interact. ICMI08 Post-Conf. Workshop*, 2008.

[141] O. Watts, J. Yamagishi, S. King, and K. Berkling, 'Synthesis of child speech with HMM adaptation and voice conversion', *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 1005–1016, 2010, doi: 10.1109/TASL.2009.2035029.

[142] K. Y. Chenpeng Du, 'Speaker Augmentation for Low Resource Speech Recognition', *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, pp. 7719–7723, 2020.

[143] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, 'Improving children's speech recognition through out-of-domain data augmentation', *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-September-2016, pp. 1598–1602, 2016, doi: 10.21437/INTERSPEECH.2016-1348.

[144] H. K. Kathania, V. Kadyan, S. R. Kadiri, and M. Kurimo, 'Data Augmentation Using Spectral Warping for Low Resource Children ASR', *J. Signal Process. Syst.*, vol. 94, no. 12, pp. 1507–1513, Dec. 2022, doi: 10.1007/S11265-022-01820-0/TABLES/6.

[145] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, 'Voice Conversion Based Data Augmentation to Improve Children's Speech Recognition in Limited Data Scenario', 2020, doi: 10.21437/Interspeech.2020-1112.

[146] Y. Zhang *et al.*, 'BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition', *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1519–1532, Oct. 2022, doi: 10.1109/JSTSP.2022.3182537.

[147] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, and T. Nakatani, 'Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders', in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6166–6170. Available: https://ieeexplore.ieee.org/abstract/document/8682890/

[148] R. Fan and A. Alwan, 'DRAFT: A Novel Framework to Reduce Domain Shifting in Self-supervised Learning and Its Application to Children's ASR.' Proc. Interspeech 2022, 4900-4904, doi: 10.21437/Interspeech.2022-11128.

[149] J. Bai *et al.*, 'JOINT UNSUPERVISED AND SUPERVISED TRAINING FOR MULTILINGUAL ASR', *ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, pp. 6402–6406, 2022, doi: 10.1109/ICASSP43922.2022.9746038.

[150] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, 'Unsupervised Speech Recognition', *Advances in Neural Information Processing Systems* 34 (2021): 27826-27839. Available: https://github.com/pytorch/fairseq/tree/

[151] A. Narayanan *et al.*, 'Toward Domain-Invariant Speech Recognition via Large Scale Training', *2018 IEEE Spok. Lang. Technol. Workshop SLT 2018 - Proc.*, pp. 441–447, Feb. 2019, doi: 10.1109/SLT.2018.8639610.

[152] F. Wu, L. Paola Garcia, D. Povey, and S. Khudanpur, 'Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network', 2019, doi: 10.21437/Interspeech.2019-2980.

[153] P. Gurunath Shivakumar and S. Narayanan, ‘End-to-end neural systems for automatic children speech recognition: An empirical study.’ *Computer Speech & Language* 72 (2022): 101289.

[154] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, ‘Kid-Whisper: Towards Bridging the Performance Gap in Automatic Speech Recognition for Children VS. Adults’. arXiv, Sep. 18, 2023. doi: 10.48550/arXiv.2309.07927.

[155] C.-L. Tai, H.-S. Lee, Y. Tsao, and H.-M. Wang, ‘Filter-based Discriminative Autoencoders for Children Speech Recognition’, *arXiv preprint arXiv:2204.00164* (2022).

[156] Z. Fan, X. Cao, G. Salvi, and T. Svendsen, ‘Using Modified Adult Speech as Data Augmentation for Child Speech Recognition’, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[157] P. N. Sudro, A. Ragni, and T. Hain, ‘Adapting pretrained models for adult to child voice conversion’, in *2023 31st European Signal Processing Conference (EUSIPCO)*, IEEE, 2023, pp. 271–275.

[158] J. J. Burred, E. Ponsot, L. Goupil, M. Liuni, and J.-J. Aucouturier, ‘CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition’, *PLOS ONE*, vol. 14, no. 4, p. e0205943, Apr. 2019, doi: 10.1371/journal.pone.0205943.

[159] L. Mcinnes and J. Healy, ‘UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction’, *arXiv preprint arXiv:1802.03426* (2018).

[160] ‘SpeD 2023 – The 12th Conference on Speech Technology and Human-Computer Dialogu’. Available: https://sped.pub.ro/

[161] M. A. Farooq, W. Yao, G. Costache and P. Corcoran, "ChildGAN: Large Scale Synthetic Child Facial Data Using Domain Adaptation in StyleGAN," in *IEEE Access*, vol. 11, pp. 108775-108791, 2023, doi: 10.1109/ACCESS.2023.3321149.

[162] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, ‘Analyzing and improving the image quality of stylegan’, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

Appendix A

A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis.

Authors: Rishabh Jain (RJ), Mariam Yiwere (MY), Dan Bigioi (DB), Peter Corcoran (PC) and Horia Cucu (HC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	RJ: 80%, DB: 10%, PC: 10%
Experiments and Implementation	RJ: 100%
Background	RJ: 90%, DB:10%
Manuscript Preparation	RJ: 75%, MY: 5%, DB: 5%, PC: 10%, HC: 5%

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 2, 2022, accepted April 20, 2022, date of publication April 28, 2022, date of current version May 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3170836

A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis

RISHABH JAIN[1], (Graduate Student Member, IEEE), MARIAM YAHAYAH YIWERE[1], DAN BIGIOI[1], (Graduate Student Member, IEEE), PETER CORCORAN[1], (Fellow, IEEE), AND HORIA CUCU[2], (Member, IEEE)

[1]School of Electrical and Electronics Engineering, National University of Ireland Galway, Galway, H91 TK33 Ireland
[2]Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, RO-060042 Bucharest, Romania

Corresponding author: Rishabh Jain (rishabh.jain@nuigalway.ie)

This work was supported by the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project (2020–2023) funded by the Disruptive Technologies Innovation Fund (DTIF), Established under Project Ireland 2040 through the Department of Enterprise, Trade, and Employment with Administrative Support from Enterprise Ireland.

ABSTRACT Speech synthesis has come a long way as current text-to-speech (TTS) models can now generate natural human-sounding speech. However, most of the TTS research focuses on using adult speech data and there has been very limited work done on child speech synthesis. This study developed and validated a training pipeline for fine-tuning state-of-the-art (SOTA) neural TTS models using child speech datasets. This approach adopts a multi-speaker TTS retuning workflow to provide a transfer-learning pipeline. A publicly available child speech dataset was cleaned to provide a smaller subset of approximately 19 hours, which formed the basis of our fine-tuning experiments. Both subjective and objective evaluations were performed using a pretrained MOSNet for objective evaluation and a novel subjective framework for mean opinion score (MOS) evaluations. Subjective evaluations achieved the MOS of 3.95 for speech intelligibility, 3.89 for voice naturalness, and 3.96 for voice consistency. Objective evaluation using a pretrained MOSNet showed a strong correlation between real and synthetic child voices. Speaker similarity was also verified by calculating the cosine similarity between the embeddings of utterances. An automatic speech recognition (ASR) model is also used to provide a word error rate (WER) comparison between the real and synthetic child voices. The final trained TTS model was able to synthesize child-like speech from reference audio samples as short as 5 seconds.

INDEX TERMS Text-to-speech, child speech synthesis, tacotron, multi-speaker TTS, alternative WaveRNN, MOSNet, subjective MOS.

I. INTRODUCTION

The bulk of recent research into human speech has focused on neural network techniques to improve speech understanding and recognition or to provide simplified, high-quality text-to-speech (TTS) models that can directly convert written text into natural speech. The most highly developed domain for such research has a focus on spoken English and is based on native-speaker adult voice data samples. Automated speech recognition (ASR) is a core element of modern consumer technology user interfaces employed in smart-speaker and voice command interfaces. For interactive chatbot and voice services, TTS models are also important, and the most advanced models can incorporate emotional and prosodic elements into the generated speech output.

More recent research into low-resource languages and other low-resource aspects of human speech, such as accented and prosody-aligned speech has started to see improvements for both ASR and TTS [1]. Another aspect of human speech of growing importance is that of child speech. Child speech differs significantly from those of adult speech, falling into a narrow range of variation and with higher pitch levels. Furthermore, children's speech patterns are more inarticulate and can vary widely in terms of volume, pacing, and emotional expressivity. These challenges are further amplified by the relatively small number of public child speech corpora that are available with useful annotations.

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

Current work done on TTS for child's voices is limited. This is mainly due to the lack of child voice datasets and difficulty in creating such datasets. As TTS models require hundreds of hours of annotated data for training [2], performing TTS for child voices can be quite challenging. The focus of this work is to explore the potential of state-of-the-art (SOTA) TTS to build a pipeline for the synthesis of children's voices with low data requirements. More specifically, if we can build such a pipeline and demonstrate that it can reliably synthesize a useful number of distinct children's voices, this pipeline would enable the creation of large synthetic datasets that could further improve other aspects of child speech research such as automatic speech recognition (ASR), speaker recognition, etc. To better elaborate on this hypothesis, it is useful to review current SOTA in TTS technologies, followed by a similar consideration for review in child speech research.

A. RELATED RESEARCH IN TTS

Early research work on TTS synthesis can be traced back to four/five decades ago when the task of TTS was commonly tackled using concatenative and parametric approaches [3]–[7]. Although these early methods were successful in generating speech from text, they generally lacked naturalness. The audio generated using these approaches was kind of muffled and sounded very robotic.

Recent state-of-the-art TTS models are largely based on deep neural networks (DNN) and can achieve more natural-sounding/human-like synthesized speech. With the introduction of Tacotron [8], a neural sequence to sequence the TTS model, the quality of speech synthesis improved significantly. While there are newer approaches that are more efficient or use smaller models, etc., it is still representative of SOTA for the quality of the synthesized speech and is used as a benchmark for comparison with newer methods. Nonetheless, Tacotron TTS is not very robust as it sometimes skips certain words and it also suffers from low inference speed [9]. Several methods have since been proposed to improve upon it such as Tacotron2 [10], FastSpeech [11], FastSpeech2 [12], Transformer TTS [13], FlowTTS [14], GlowTTS [15], etc. Similarly, there have been several improvements over the quality of synthesized waveforms by the introduction of SOTA Vocoders such as WaveNet [16], WaveGlow [17], MelGAN [18], Hifi-Gan [19], WaveRNN [20], etc. These TTS models supported single speaker synthesis, but Deepvoice2 [21], introduced the use of speaker verification models [22]–[25] to achieve Multi-speaker TTS [26]–[34].

B. CHILD SPEECH – LITERATURE AND CHALLENGES

While all SOTA TTS systems rely on large datasets to train, the datasets mostly comprise speech taken from adult native English speakers; hence, for low-resource languages and other target groups such as non-native adult speakers and child speakers, there remain challenges developing effective and suitable TTS models. Specifically, in comparison with adult TTS, child TTS has gained very little to no attention from the TTS research community. With the current trend of data-hungry DNN-based TTS, TTS for children has practically been neglected due to the lack of large publicly available children's speech datasets suitable for training such networks. Prior to this DNN era, researchers worked on TTS for children using HMM-based models [3], [6].

Collecting data for child speech research can be a challenging task. Most TTS datasets are created in studios with expensive equipment: an adult will be using a microphone to create a clean, noiseless, easy to understand, and meaningful audio. This task is not easy to produce and even more difficult to implement with a child.

One of the main differences between adult speech and child speech is the fundamental frequency. The pitch for children is significantly higher than that of an adult [35]–[38]. The pitch for an adult voice lies between 70 to 250 Hz whereas the pitch for the children's speech is between 200 to 500 Hz [39]. There is also a difference in the speaking rate of children. It was noticed that average phoneme duration is longer in children, therefore, leading to longer speaking rates as compared to adult speech [38], [40]–[42]. The vocal tract of an adult is larger as compared to children's vocal tract and therefore produces different prosody features as compared to an adult voice [43], [44]. Hence, a substantial difference in children's voice characteristics and features can be seen as compared to an adult voice.

Our work aims to solve the problem of TTS for children using DNNs. To solve this problem, the huge challenge of limited publicly available children's speech datasets must first be overcome. To this end, this study considered the use of an existing multi-speaker children's speech dataset [45], which comes with an incomplete set of utterance transcriptions. In addition, this dataset has a lot of unusable data, such as empty/blank entries, extremely long entries as well as inaccurate transcriptions. Firstly, the dataset is cleaned up to create a subset that is suitable for training a neural TTS model. Secondly, with the cleaned-up dataset, a multi-speaker TTS model is trained to generate synthetic speech for multiple child speakers as a proof of concept for children's TTS. The training involved fine-tuning an existing adult multi-speaker TTS model [33] by way of transfer learning, with a few modifications as explained in later sections. This approach involves the training of a separate speaker verification model, and it was preferred because it reduces the problem at hand in two ways:

1) To train the speaker verification network, transcriptions for the speech dataset are not required. Only the speaker identities for the utterances are needed and it can also be trained on noisy speech without any negative effects. This means that even the noisy children's speech dataset, which has incomplete transcriptions, can be useful in training the verification model.
2) Being a transfer learning process, the pretrained TTS model can be finetuned sufficiently using the resulting cleaned set of children's speech data.

Subjective and Objective Evaluation performed on the synthesized child voices confirms that the child voices generated

TABLE 1. Dataset used in this work.

Dataset	# of speakers	# of hours	# of utterances
MyST	1371	393	228,874
VoxCeleb1	1251	352	153,516
LibriSpeech	2484	1000	-
VCTK	110	44	400 each

synthetically are very close to the real child voice in terms of different acoustic features and MOS.

The rest of this paper is organized as follows. Section II describes the methodology and datasets used in this study. The experiments are presented in Section III, the result and evaluation in Section IV, and finally, the conclusion and future work in Section V.

II. PROPOSED METHODOLOGY

A. DATASETS USED IN THIS STUDY

The nature of this study, considering the challenge of limited children's speech datasets and the multi-step training process involved, calls for the use of multiple large datasets, including adult speech datasets. All these datasets are described in Table 1.

- **MyST** [45]**:** My Science Tutor (MyST) children's corpus consists of child speech collected using the interaction of the student with a virtual science tutor. The data consists of 393 hours of child speech collected from 1371 students producing a total of 228,874 utterances. 45% of the data is transcribed at word-level leading to about 103,082 utterances, around 208 hours presented in a *.trn* file format. The MyST corpus is used for this paper because it is the biggest corpus of child speech freely available for research use.
- **VoxCeleb1** [46] **:** VoxCeleb 1 contains audio recordings of celebrity voices extracted from YouTube. It contains 153,516 utterances from 1,251 speakers.
- **LibriSpeech** [47] **:** LibriSpeech is a read English speech dataset derived from audiobooks. The data contains approximately 1000 hours of adult speech data from 2400 speakers. The data is divided into two sets, "clean" and "other" where the clean set contains less noisy data as compared to the other set. The "clean" set contains 460 hours of data, and the "other" set contains 540 hours of data.
- **VCTK** [48] **:** This dataset contains speech recordings from 110 English speakers each reading about 400 sentences from a newspaper. The data contains recordings from various English accents and is highly used in multi-speaker TTS research.

1) PROBLEMS IDENTIFIED IN MYST DATASET

A study on the MyST dataset was performed to measure the amount of data in MyST with and without a transcript. This was done to extract data available with annotation and to see if it can be used for training TTS. A comparison between the complete MyST dataset and filtered MyST dataset where transcripts are available is presented in Table 2. This table provides information on the utterance count and duration of utterance concerning the duration range.

TABLE 2. MyST dataset comparison [complete vs with transcript].

	MyST (Complete)		MyST (with transcripts)	
Seconds (range)	# of utterances	Duration (in hours)	# of utterances	Duration (in hours)
0-5	113,219	62.19	51,350	27.98
5-10	43,782	87.78	20,723	41.78
10-15	**22,321**	**75.80**	**11,096**	**37.78**
15-20	11,477	54.86	6,067	29.04
20-30	9,282	61.89	4,991	33.28
30-40	2,796	26.50	1,542	14.61
40-50	930	11.40	517	6.32
50-60	347	5.24	184	2.78
60-70	146	2.61	83	1.48
70-80	74	1.52	38	0.78
80-90	53	1.25	23	0.54
90-100	19	0.49	5	0.13
100 Above	41	1.95	17	0.94
Total	**228,874**	**393.51**	**103,082**	**197.48**

From Table 2, it was observed that 197.5 hours of child speech data is available with annotation. Although a lot of this data can't be used having different memory requirements on different GPUs. In our experiments, that data between the range of 10-15 seconds to be most useful.

Some initial experiments were performed on the MyST dataset without using the Multi-speaker TTS approach (see section III.A). The results obtained from these experiments were unintelligible. The output waveforms did not have any phonetic meaning and were missing quite some pronunciations. On a more detailed manual inspection of the MyST dataset, a few common problems were identified. The transcripts of some example audio files are listed below to illustrate the problems in the MyST dataset:

- Audio files containing noise in their utterances without any phonetic meaning.
  - "<noise>"
  - "it's glowing <breath>"
- Audio files that are not coherent or indiscernible.
  - "in oxygen right <indiscernible>"
  - "can hear sound because of that <indiscernible>"
- Audio files are too small in length
  - "energy <noise>"
- Audio files are too long
  - "it's trying to show us that all the things that it needs all the things that the plants needs to grow it needs soil on the bottom it needs at least a ground a top the a a top to lay on for the plant to grow so you can see it that's only with flowers and plants it's not with vegetables and it needs and it needs the energy from the sunlight to grow and it needs water because somebody's watering the plant."

TABLE 3. MyST vs TinyMyST.

	MyST (Complete)	MyST (with transcripts)	TinyMyST (Usable for TTS)
Speakers	1371	738	670
Duration (in hrs)	393.5	197.5	19.22
# of utterances	228,874	103,082	7152
Mean duration per speaker	17.22 mins	16.05 mins	1.73 mins
Speaker with most data	013023 (110.38 mins)	013023 (81.71 mins)	013023 (8.77 mins)
Speaker with least data	012002 (0.96 secs)	007389 (0.96 secs)	018216 (10.2 secs)

mins: minutes, secs: seconds

- Transcription containing text with no phonetic information.
  - "(()) (()) (())"
- Repetition of words/stammering noticed in children's voices.
  - "um we measured how big a millimeter meter is a meter and a kolome- a ∗ kilometer ∗"

Our examination of MyST led us to further clean the MyST dataset for TTS training. In this process a subset of MyST, hereafter referred to as TinyMyST was created.

2) TINYMYST

It is a small subset of the MyST dataset created using various pre-processing scripts to make the data suitable for TTS acoustic model training. MyST was cleaned to select only audio files with existing transcriptions. All audio files lesser than 10 seconds and greater than 15 seconds were removed. The utterances shorter than 10 seconds contained mostly noise or unintelligible speech and those longer than 15 seconds were removed to avoid GPU memory overflow during training. All the transcript files were converted from *.trn* format to *.txt* file format.

The TinyMyST dataset still contains a lot of noisy data. Some of the excessively noisy data were removed manually by inspecting the transcripts and listening to the audio samples. The data obtained after cleaning contained 7152 utterances and accounted for 19.22 hours. A detailed comparison of MyST and TinyMyST datasets was performed to see differences in the two datasets in terms of speaker identities and utterances (see Table 3). TinyMyST dataset on average contained 1.72 minutes per speaker having 670 speakers. Speaker identity '013023' had the most data with 8.77 minutes and speaker identity '018216' has the least data with 10.01 seconds. The speaker '013023' had the most data in MyST as well to be around 110 minutes.

To extract more TTS usable data, an audio sample from more than 15 seconds long can be used to split them into smaller chunks. A forced aligner[1] is used to align the audio files with transcripts. Time alignment information from the alignments to split the longer audio files into smaller samples, however, it was observed that the audio alignment was not very accurate for the child speech and there were a lot of mismatches between the transcripts and audio files. This was probably due to fact that the pretrained forced aligner was trained on adult speech and doesn't work very well for aligning child speech. Therefore, TinyMyST was used (as described earlier) for performing all the child TTS experiments.

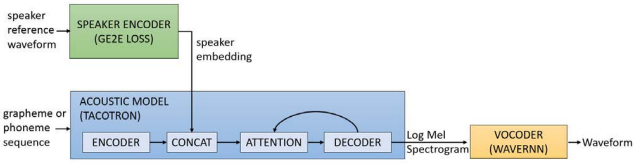


FIGURE 1. Model Overview: Speaker Encoder, Acoustic Model, and Vocoder Models trained independently (from [33]).

3) DATA PREPROCESSING FOR TTS USAGE

LibriSpeech and TinyMyST datasets were preprocessed as per the guidelines mentioned in LibriTTS [49]. The LibriTTS dataset was specifically created for TTS research, therefore similar guidelines were followed in our experiments. The following changes were made:

- Audio files were converted to 16-bit depth audio files with 24Khz sampling rate (WAV format), This was done using the pydub[2] audio library.
- Text data was normalized by replacing abbreviations and punctuations.
- Whitespaces were normalized
- All characters were made uppercase.

B. MULTI-SPEAKER TTS MODEL

The neural network used to achieve TTS for children in this study is based on [33], It works by combining a speaker verification network with the SOTA Tacotron TTS model. Though Tacotron is SOTA for TTS, it was designed to be trained using a single-speaker speech dataset such as the LJSpeech [50] dataset, hence, it can only synthesize speech with acoustic characteristics of the single speaker whose data was used in training. To function effectively for multiple speakers, Tacotron needs to be adapted for that purpose. This adaptation has been achieved in this multi-speaker TTS model [33] by introducing different speaker identities in the form of speaker embeddings as additional input to the Tacotron network. As a result, the multi-speaker TTS [33] comprises three different neural network models, each of which focuses on a specific subtask namely, Speaker Encoder used for speaker verification task, Acoustic model used for spectrogram synthesis, and a Vocoder for audio waveform generation (as shown in Figure 1).

For our work, generalized end-to-end (GE2E) loss was used for speaker verification [22], Tacotron1 as an acoustic model [8], and WaveRNN as Vocoder [20]. The original approach [33] is adapted for child speech synthesis by first

[1] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[2] https://github.com/jiaaro/pydub

pretraining the model on an adult speech dataset after which, it is fine-tuned with the child speech dataset.

The speaker encoder generates speaker embeddings, encoding speaker identity information extracted from the utterances. Similar voices are mapped closer to each other in a latent space representation. The acoustic model generates spectrograms from text conditioned on the speaker embeddings. The vocoder then converts these spectrograms into audio waveforms. At inference time, a short reference utterance (ground truth) of a child's voice is passed through the speaker encoder to generate the corresponding speaker embeddings, on which the acoustic model will be conditioned. The three different neural network models are described as follows.

1) SPEAKER ENCODER

The first stage of the multi-speaker TTS training involves the training of a speaker verification (speaker encoder) model. Speaker Verification is the process of determining if an utterance belongs to a specific speaker. The speaker encoder is used to train the model for the speaker verification task using a mix of noisy and clean speech data without transcripts. The data used consists of both adult and child speech data from thousands of speakers (see Table 4). This was done to introduce both child and adult speakers in the model for better generalization. The output of this model conditions the acoustic model to generate the required mel-spectrograms from a reference speech signal of the target speaker. The model is trained to capture the characteristic features of different speakers.

The model takes input as log mel-spectrograms computed from utterances of each speaker, trains using the GE2E loss and converts them into a fixed dimensional vector called d-vectors. These d-vectors are optimized over GE2E loss to differentiate the speakers, such that the same speakers have embeddings with high cosine similarity and different speakers are far apart in the embedding space.

During training, complete utterances are segmented into partial utterances of 1.6 seconds. These parameters were kept the same as explained by authors [51], [22]. The utterance embedding is calculated using 800ms windows for inference, with a 50% overlap. The silence was removed from the utterances using the webrtcvad[3] tool for Voice Activity Detection (VAD). Each segment is passed through the network individually, the outputs are averaged and normalized to create the final utterance embedding as described in [22].

The encoder model is trained using 4 datasets, MyST, VoxCeleb1, LibriSpeech, and VCTK. Equal Error Rate (EER) is used as a metric for the validation of the speaker encoder. The default EER metric from [51] is used in this work as authors of [33] have not explicitly specified the training, test, and validation criterion they are using for EER calculation. The EER values are presented in Table 4. The model trained for one million steps was used in the multi-speaker TTS model as relatively insignificant improvements were seen in the EER after this point.

[3]https://github.com/wiseman/py-webrtcvad

TABLE 4. Speaker encoder training details.

Dataset used for Encoder Training	Size (in hours)	Iterations	EER
- MyST - VCTK - VoxCeleb1 - LibriSpeech [Other]	1329	1M	5%

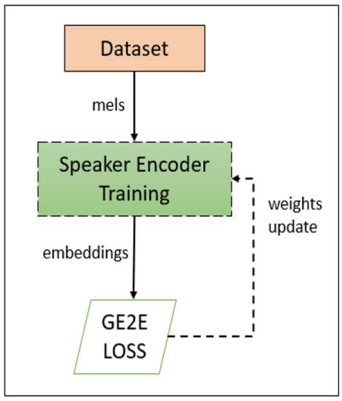


FIGURE 2. Pipeline for Speaker Encoder training. The dotted line represents the training loop for the Speaker Encoder training.

All the datasets were pre-processed into the coding format required for training the encoder as described in [51]. Even though half the MyST dataset is not transcribed, the complete MyST dataset can be used for Speaker Encoder training as it does not require any transcription data. The pipeline for the speaker encoder training can be seen in Figure 2.

A UMAP projection [52] is created to visualize the training by taking a random set of 10 utterances from 10 speakers. Utterances with similar embeddings are located close to each other in the latent space representation and have similar speaker characteristics.

This model creates individual clusters of speaker embeddings as can be seen in the UMAP projection (see Figure 3. Each point on UMAP represents an utterance. The same color points represent the same speaker. Encoder gradually learns to separate the speakers. Initially, there is a lot of overlap across speakers, but eventually, each speaker has their utterances clustered and well separated from the other speakers. The training evolves with increased training steps.

2) TACOTRON ACOUSTIC MODEL

For the speech spectrogram synthesis, the TTS model architecture and hyperparameters used in this study are the same as in the work of [51] (More details are provided in Section III). The authors used a modified version of the original Tacotron architecture [8]. The model consists of an encoder, an attention-based decoder, and a post-processing network. Since Tacotron is originally a single-speaker TTS model, it was modified to work for multi-speaker TTS by connecting the speaker encoder to it. Speaker embeddings from the encoder are concatenated with text (character/phoneme)

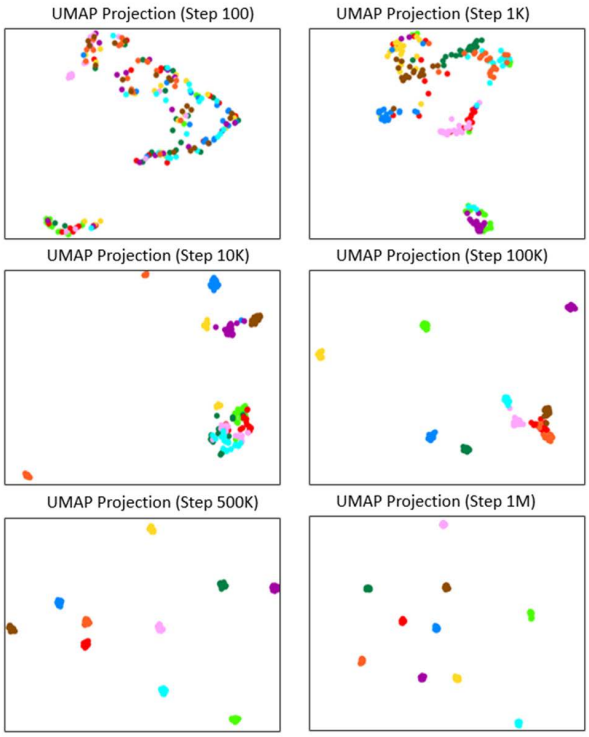
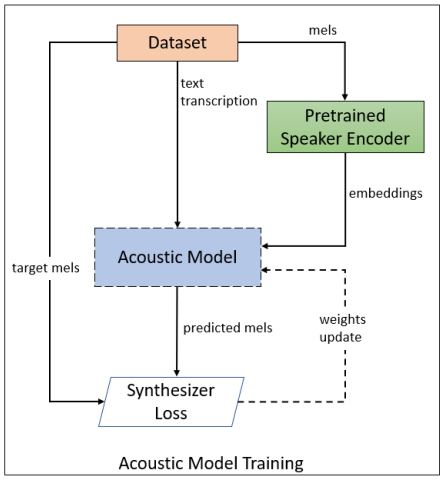
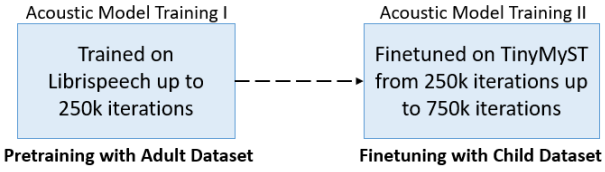


FIGURE 3. UMAP projections at different training steps for speaker encoder training. Ten different colors represent ten different speakers with ten utterances each.



(a)



(b)

FIGURE 4. Pipeline for Acoustic Model training. A model with solid contour represents the pretrained model. Dotted contours represent the acoustic model training loop. Fine-tuning step for Acoustic model training. Acoustic model training I represent the acoustic model being trained with LibriSpeech dataset for up to 250k iterations. Acoustic model training II represents fine-tuning the acoustic model I with the TinyMyST dataset from 250k iteration onwards up to 750k iteration.

embeddings from the text encoder, after which an attention mechanism is applied prior to decoding into a spectrogram. Unlike the speaker encoder, the acoustic model takes in both audio(utterance) and associated text(transcript) as inputs.

In this work, the acoustic model was first trained with only adult speech data (acoustic model training I), specifically, the Librispeech 'clean' data, until it started to converge at 250k steps and then finetuned with the TinyMyST child speech dataset (acoustic model training II) for up to 750k additional steps (more details in Section III). The pipeline for the acoustic model training can be seen in Figure 4.

3) WAVERNN VOCODER

The vocoder used is WaveRNN [20], which is an improvement over the WaveNet [16] originally used by the authors of [33]. WaveRNN is a recurrent network for performing sequential modeling of audio from mel-spectrograms. An alternative version of WaveRNN is used, having a few architectural changes as provided by the author in [53] due to the popularity of the model as it reduces sampling time while maintaining high output quality. WaveRNN uses Gated Recurrent Unit (GRU) in comparison to convolutions used in WaveNet. The input mel-spectrograms and their corresponding waveforms are segmented at each timestamp. A 1D Resnet-like model is used to generate features for layered connections in the alternative WaveRNN architecture. The upsampling is also performed on the mel-spectrogram to match the length of the target waveform. The resulting vector is passed through a combination of GRU and dense layer transformations in four-way connections. These connections are concatenated at different steps to generate the corresponding vector representation. This vector is passed through two dense layer connections which finally generate the encoding of raw audio. The output audio is generated at a 16-bit depth and 16 khz sampling rate.

The predicted mels from the acoustic model trained on LibriSpeech (from acoustic model training I) were used to train the vocoder. The vocoder trained up to 250k iterations was used to generate all waveforms in this study. The pipeline for vocoder training can be seen in Figure 5. Fine-tuning experiments with the TinyMyST dataset didn't improve the quality of the vocoder (more discussion in Future work).

Vocoder for child TTS hasn't been explored before in detail. This is a new area of research. It was observed that WaveRNN has popularly been used as a universal vocoder [54]–[56] and it evidently works well with unseen speakers in multi-speaker models as well [57]. Therefore, for the scope of this paper, WaveRNN (trained on LibriSpeech) is used as a universal vocoder with synthetic child voices.

III. EXPERIMENTS

A. INITIAL EXPERIMENTS

In our initial experiments, multiple SOTA TTS models [10], [13]–[15], [21] were unsuccessfully trained, including

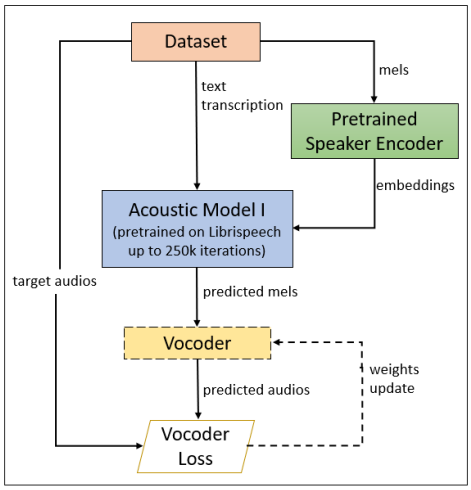


FIGURE 5. Pipeline for Vocoder training. Models with solid contours are pretrained models. The dotted contour represents the training loop for Vocoder.

Tacotron 2,[4] using the transcribed subset of the MyST dataset. Figure 6 shows an example of an alignment plot from Tacotron 2 training. As can be seen, there was no sign of alignment even after 200k iterations.

Further experiments were conducted using the cleaned subset of MyST (TinyMyST), which showed some alignments as seen in the Tacotron 2 alignment plot in Figure 7. However, though child-like in terms of pitch, the synthesized speech signals were completely unintelligible. Missing information such as 'End of sentence' was observed which mostly contained noise content.

Next, fine-tuning the pretrained NVIDIA Tacotron 2 model on a single child's MyST dataset utterances resulted in slightly intelligible but highly robotic and unnatural synthesized speech. Figure 8 shows the improved alignment plot from the finetuned Tacotron2.

Since there were not enough MyST utterances for a single child to sufficiently train Tacotron2, different multi-speaker TTS models were explored [21], [26], [27], [58] and the speaker verification-based method [33] produced the most promising results. Hence, this method was used in our main experiments.

B. MAIN EXPERIMENTS

As seen in our methodology (Section II.B), a modified approach based on [33] was used by incorporating an extra layer of fine-tuning in the training step.

The proposed neural child voice TTS was trained on a Tesla V100 GPU. Each of the three networks – Speaker Encoder, Acoustic model, and Vocoder were trained separately.

The Speaker Encoder was trained with a batch size of 128 and a learning rate of 0.0001. The model was trained for 15 days for up to 1M steps. EER of 5% was observed at this point with no further improvement afterward.

[4]NVIDIA/tacotron2: https://github.com/NVIDIA/tacotron2

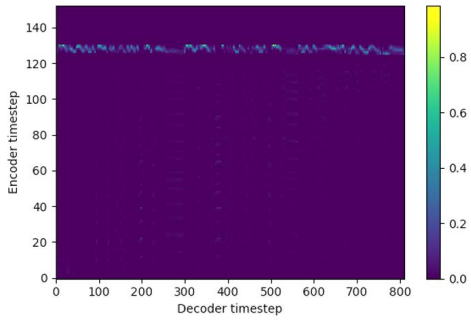


FIGURE 6. Alignment plot for Tacotron 2 trained with MyST dataset for up to 200k steps.

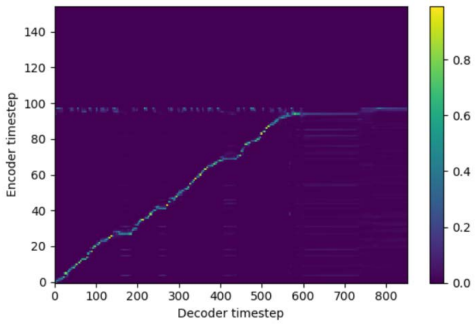


FIGURE 7. Alignment plot for Tacotron 2 trained up to 200k steps with TinyMyST Dataset.

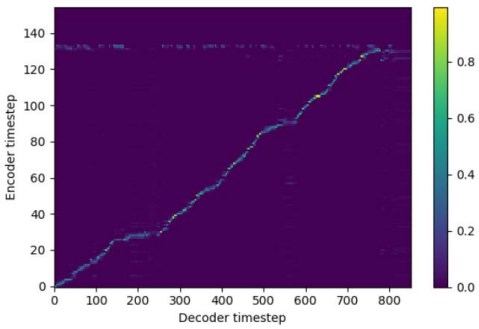


FIGURE 8. Alignment plot for Tacotron 2 trained up to 200k steps with TinyMyST Dataset, pretrained with LJ Speech Dataset up to 100k steps.

Additional parameters settings are mentioned here.[5] The default embedding size of 256 was used for this training.

For the acoustic model, the network was trained using a learning rate of 0.0001 for 250K steps (pretraining) and 0.00001 for 750k steps (fine-tuning). The batch size was kept constant at 72. Entire training (up to 750k steps) took 9 days to complete. Additional parameters details were kept the same as Tacotron 1, these details are mentioned here.[6]

The alignments plot for encoder-decoder timestamps can be seen in Figure 9, the x-axis represents the encoder timesteps and the y-axis represents the decoder timesteps of Tacotron training. The training is done on LibriSpeech up to 250k steps generated a good alignment plot. Alignment weakens when switched to TinyMyST Dataset, but it

[5]Encoder Hyperparameters: https://github.com/CorentinJ/Real-Time-Voice-Cloning/blob/master/encoder/params_model.py

[6]Acoustic Model Hyperparameters: https://github.com/CorentinJ/Real-Time-Voice-Cloning/blob/master/synthesizer/hparams.py

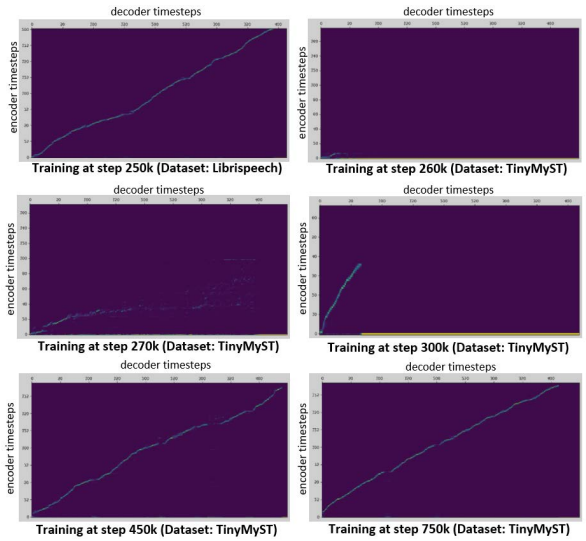


FIGURE 9. Alignment plots at different training steps during transfer learning from adult to child Tacotron TTS.

gradually improves with increasing training steps. During inference, our model was tested on multiple checkpoints taken at intervals of 50k iterations. A few of these iterations are mentioned in Figure 9. Even though the alignment at some of these steps looks the same, an improvement was noted over time with the synthesized child voices. This was determined subjectively during training by listening to the synthetic child speech generated. The training was halted at 750k steps as improvements in the alignment graph had become imperceptible after 700k steps. The output waveform did not show any improvements beyond this step. The model trained up to 750k iterations is used to provide audio samples in this paper.

The Vocoder was trained at a batch size of 128 and learning rate of 0.0001 and took 4 days of training to reach 250k iterations. Most of the parameters for the vocoder were kept the same as the original code.[7]

The synthetic child voices during inference were natural sounding and the trained model demonstrated an ability to synthesize quite challenging phrases that were unseen in the TinyMyST dataset. This was tested by using 'tongue twisters' as a reference text for synthesizing speech. However, it was also noted that some phonemes were not synthesized correctly and lost their meaning during synthesis. These findings are discussed in more detail in Section IV.

Code-related material and synthesized speech from these experiments will be made available in our GitHub Repository.[8]

IV. RESULTS AND EVALUATION

The evaluation in TTS is usually done by taking a Mean Opinion Score (MOS) [59] on the synthetic speech for Speech Similarity and Speech Naturalness.

[7]Vocoder Hyperparameters: https://github.com/CorentinJ/Real-Time-Voice-Cloning/blob/master/vocoder/hparams.py

[8]GitHub for this paper: https://github.com/C3Imaging/ChildTTS

There are many objective and subjective evaluation methods proposed by researchers [60]–[66]. These traditional speech evaluation methods work well for evaluating adult speech but are not so suitable for child speech. A perfect adult speech will contain fluent pronunciation of a word/phoneme however this is not the case for most child speech. Naturalness in child speech includes pauses, breaks, and pronunciation difficulties in the speech. Other challenges were noted with the start and end of phrases where children tend to be somewhat hesitant when starting a phrase and may wander towards the end of one. Children can also mispronounce words, or struggle with the phonetics of a particular phrase.

These characteristics tend to manifest in the speech model and a range of artifacts were noted that affect the quality of the phrases synthesized by our pipeline. It was noted that the first or last words in many phrases were either missed entirely or subject to various distortions or artifacts. In the middle of a phrase, there could occasionally be slurring or arbitrary elongating of one or more words. Another artifact observed was that the pace or tone of voice could change abruptly in the middle of a phrase. Despite these artifacts, the majority of phrases were quite intelligible, and a large proportion was also very natural sounding. Therefore, there is a need for a better subjective evaluation method for child speech synthesis.

In the following, we present the results obtained using the proposed subjective evaluation method and the various listening tests performed (subsection 4.A), two objective evaluation methods, based on MOSNet (subsection 4.B) and an ASR system (subsection 4.D) and we evaluate the similarity of the synthesized speech and natural child speech (in subsection 4.C).

A. PROPOSED SUBJECTIVE EVALUATION METHOD

To check the phonetic coverage of our child speech TTS, Harvard sentences [60] were used, which are a set of 720 phonetically balanced sentences. These sentences cover most of the phoneme range and were designed to be implemented with Voice over Internet Protocol (VoIP) technology. These texts were used to generate synthetic child speech. This was done to check the subjective quality of synthesized audio with respect to phoneme coverage.

Our evaluation method uses a MOS-like evaluation with different categories for scoring. When generating synthetic voices using Harvard sentences, it was observed that some sets of phonemes were not pronounced correctly even when synthesized using different reference child speakers (more detail in a later section). After our initial subjective study of these 720 synthesized audio samples, it was decided that a more detailed evaluation protocol was required to address the various artifacts observed and identify what additional data samples might be needed to further improve our model. For this reason, our evaluation was performed in two phases. For each of the two phases, different evaluators were gathered to perform the speech evaluation. Each evaluator was asked to listen to synthetic audio files using Headphones/Earphones

in a noise-free environment. They were asked to rate each of the synthetic voices assigned to them from a range of 1 to 5, for each of the different categories in two phases. The categories included Speech Intelligibility, Voice Naturalness, and Voice Consistency. Voice Consistency contained three sub-categories of its own namely, Start of Phrase quality, Middle of Phrase quality, and End of Phrase Quality.

Evaluation data was provided in a OneDrive Environment. All the synthetic voices were shared in a common OneDrive folder to the evaluators and a common spreadsheet was circulated containing the utterance ID of Harvard sentences used for synthesizing a child's voice. While listening to many different natural child voices, it was also noticed that recorded child audio can be a difficult task to understand if not provided with a suitable transcript. Some of the child's speech can be non-meaningful as mentioned in problems with the MyST section. After performing many different tests and trials using child speech, the use of transcripts as a part of MOS-based evaluation is considered to be a more natural way of evaluating child speech. Therefore, corresponding transcript information is also provided in the spreadsheet to each evaluator to base their conclusion on 'what they hear in child audio' and 'what they read in child transcripts'. This way more coherent patterns can be observed among the phonemes and graphemes in a child's voice for each of the mentioned categories. An example of this spreadsheet can be seen here.[9]

By performing the evaluation using OneDrive environment, it was easy to distribute the synthetic speech files to different evaluators without having to spend time and resources on expensive Mushra-based evaluations [67] or crowdsourcing the evaluation task on platforms like Amazon Mechanical Turk (AMT) [59], [68]. Mushra-based evaluations were also avoided due to potential biases that can occur in these tests and how these biases can impact synthetic child voice evaluation for MOS [69]. Most of these TTS evaluations have been conducted before with synthetic adult speech, this novel synthetic evaluation is implemented for first-time with synthetic child speech. Using a common spreadsheet made it effective to perform analysis of spreadsheet for MOS using pandas and other python-based tools.

1) PHASE-I EVALUATION

For the phase-I evaluation, all 720 Harvard sentences were generated using our proposed TTS method. Two random reference utterances were selected from the TinyMyST dataset and were used to generate all the Harvard sentences. These 720 sentences were shared among 5 evaluators in a spreadsheet document, who rated the voices from 1 to 5 based on **Speech Intelligibility** and **Voice Naturalness**. The MOS ratings from 1 to 5 were further explained in the spreadsheet file as can be seen in Table 5.

[9]Example Spreadsheet: https://www.github.com/C3Imaging/ChildTTS/blob/main/synthetic%20evaluation%20example/Example.xlsx

TABLE 5. MOS (from 1 to 5) explained for speech intelligibility and voice naturalness.

Score	Speech Intelligibility	Voice Naturalness
(5)	Voice is clear, all words identifiable	Voice consistently paced with similar timbre across the entire phrase; good voice quality
(4)	Voice is mostly clear; single word unclear	Some disjointness in terms of pacing/timbre; mediocre but plausible voice quality
(3)	Voice understandable; multiple words unclear	Significant differences in pacing/timbre across different portions of the phrase; weak voice consistency across different parts of the phrase
(2)	Difficult to understand most words in a phrase	Substantial differences in pacing/timbre across different portions of the phrase; distinctly different voices for different words/parts of the phrase
(1)	Difficult to understand any words in a phrase	No consistency in terms of voice or pacing across the phrase

TABLE 6. MOS from phase-I evaluation with 95% confidence interval

Categories	MOS
Voice Naturalness	**3.88±0.27**
Speech Intelligibility	**4.13±0.34**

The spreadsheet was later analyzed to get the final mean opinion score in each category. MOS of 3.88 for voice naturalness and 4.13 for speech intelligibility was observed as seen in Table 6.

An average score for each of the 720 sentences was calculated for the combined value of speech intelligibility and voice naturalness. All the 720 sentences were sorted into difficult and easy sentences with respect to the children's linguistic capabilities. This was done to keep track of Harvard sentences where synthesized speech becomes unintelligible and inarticulate.

2) PHASE-II EVALUATION

After our phase-I evaluation, a common set of sentences were observed where pronunciation sounds unintelligible at the start, middle, or end of sentences for specific words/phonemes. There was an inconsistency in voice quality. These sets of sentences are the ones that were not learned properly during training or were missing in the training dataset for child audio. To make a note of these sentences, extra categories of 'Voice Consistency' were added to the phase-I evaluation. Therefore, all the 3 sub-categories under **Voice Consistency** were used in the second phase of the evaluation. These subcategories included 'Start of Phrase Quality', 'Middle of phrase quality' and 'End of phrase quality'. The MOS ratings from 1 to 5 for each of these categories were also explained in the spreadsheet as mentioned in Table 7.

For the second phase of evaluation, the evaluation was undertaken by 20 evaluators divided into 4 groups. This was done as per the guidelines mentioned in [70] for performing MOS evaluations. For each group, a speaker identity was selected from the TinyMyST dataset. All the speaker identities were sorted, and the top 20 speaker identities were

TABLE 7. MOS (from 1 to 5) explained for voice consistency and its three sub-categories.

Score	Voice Consistency		
	Start of phrase & first word quality	Middle of phrase & central word quality	End of phrase & last word quality
(5)	First word is clear; excellent starting quality & intelligibility	Middle of phrase is clear; excellent voice quality & intelligibility	Last word has excellent quality
(4)	Understandable; minor distortions or noise; low intensity starts of first word	Understandable; minor distortions or noise; some pacing variations or slurring of single word	Understandable; minor distortions or noise
(3)	Start of first word unclear or more significant distortions of first work or start of phrase	Significant distortions of middle phrase or slurring of multiple words but still intelligible	Understandable; but significant distortions or noise
(2)	First word missing or strongly distorted; distortions or noise impact strongly on phrase intelligibility	Substantial distortions; multi-word slurring or noise impact strongly on phrase intelligibility	Not understandable; substantial distortions, missing or unintelligible words or noise
(1)	Start of phrase unintelligible, missing or severely distorted	Middle of phrase unintelligible, missing or severely distorted	Multiple words not understandable

TABLE 8. Selected speaker identity information in TinyMyST VS TTS utterances for the same speakers.

Speaker ID	Minutes	Hours	# of Real utterances	# of generated TTS utterances
002113	7.75	0.13	41	50
008045	7.88	0.13	42	50
013020	8.77	0.15	47	50
995737	5.70	0.10	30	50

selected, having the most minutes. Among these 20 identities, 4 speaker identities were randomly selected. All the 4 groups are named as '013020', '008045', '002113', and '995737', corresponding to each identity label. This approach was taken to select speakers with the most data and also to keep the process randomized. More information on these selected speakers can be seen in Table 8. This table is also used for speaker similarity and objective intelligibility experiments in the future sections.

A reference child utterance was selected randomly from each of these groups, and 50 Harvard sentences were selected randomly for each of the groups. Therefore, 50 Synthetic utterances were generated, and all the evaluators were asked to rate the utterances assigned to them.

MOS results from the phase-II evaluation are presented in Table 9. MOS of **3.95** was observed for **Speech Intelligibility**, **3.89** for **Voice Naturalness**, and **3.96** for overall **Voice Consistency** (including the three sub-categories). MOS of 4.07 was observed for 'Start of phrase quality', 4.18 for 'Middle of phrase quality', and 3.62 for 'End of phrase quality'. The MOS score implies that the quality of synthesized child speech is quite good. However, there is still

TABLE 9. MOS from phase-II evaluation with 95% confidence interval.

	SI	VN	VC		
			SP	MP	EP
013020	4.03	4.06	4.44	4.35	3.61
008045	3.70	3.63	3.94	3.92	3.20
002113	3.62	3.49	3.62	3.89	3.48
995737					
Overall MOS	**3.95±0.30**	**3.89±0.32**	4.07±0.36	4.18±0.21	3.62±0.45
			3.96±0.32		

SI: Speech Intelligibility, VN: Voice Naturalness, VC: Voice Consistency, SP: Start of phrase and first word quality, MP: Middle of phrase and central word quality, EP: End of phrase and last word quality.

room for improvement in the 'End of phrase quality' of Harvard sentences. There is information loss observed at the end of most sentences containing inarticulate and unintelligent information or noise. The reason for this information loss can be redirected back to the child dataset used for training. Even though TinyMyST is much cleaner than the MyST dataset, it still contains some of the problems seen in Section II.A.1. The information obtained from voice consistency will be discussed more in future work.

A similar experiment was also performed using the real utterances from Table 8 to obtain a baseline MOS on natural child speech. 15 random real utterances were selected from the real speakers mentioned in Table 8. Evaluators were asked to perform a similar evaluation as done in phase-II evaluation for all the selected 60 utterances. A comparison between the baseline MOS on Natural MyST and synthetically generated utterances is mentioned in Table 10.

Synthetic Speech MOS for three categories is very close to Natural Speech MOS. There is a **MOS difference** of '**0.26**' for Speech Intelligibility, '**0.16**' for Voice Naturalness, and '**0.12**' for Voice Consistency between natural and synthetic speech. From Table 10, it can be concluded that the MOS for Natural and Synthetic child speech are quite close to each other. This subjective evaluation approach is proposed as a part of this paper. Due to very limited work done on child speech synthesis, we did not find any reliable way of performing subjective evaluation over the synthesized child speech. From our experience with the evaluation of synthetic child speech, this new metric of evaluation can help evaluate synthetic child speech and can help further this area of research. It is also intended to use this proposed approach for our future work with child speech synthesis.

B. OBJECTIVE NATURALNESS EVALUATION USING A PRETRAINED MOSNET

For this objective evaluation, a pretrained MOSNet was used, which is trained on VCC 2018 dataset from Blizzard Challenge [66] comprising of adult speech. According to their paper, MOSNet predictions yield a high correlation to human ratings. As MOSNet was trained on adult speech, it is unlikely that it will generalize well for child speech. It won't be possible to train a MOSNet with child voices as there is not

TABLE 10. MOS natural speech VS MOS synthetic speech with 95% confidence interval.

	SI	VN	VC		
			SP	MP	EP
Natural Speech MOS (from MyST)	**4.21±0.42**	**4.05±0.34**	4.32±0.42	4.01±0.6	3.9±0.62
			4.08 ± 0.54		
Synthetic Speech MOS (from Table 9)	**3.95±0.30**	**3.89±0.32**	**3.96 ± 0.32**		

SI: Speech Intelligibility, VN: Voice Naturalness, VC: Voice Consistency, SP: Start of phrase and first word quality, MP: Middle of phrase and central word quality, EP: End of phrase and last word quality.

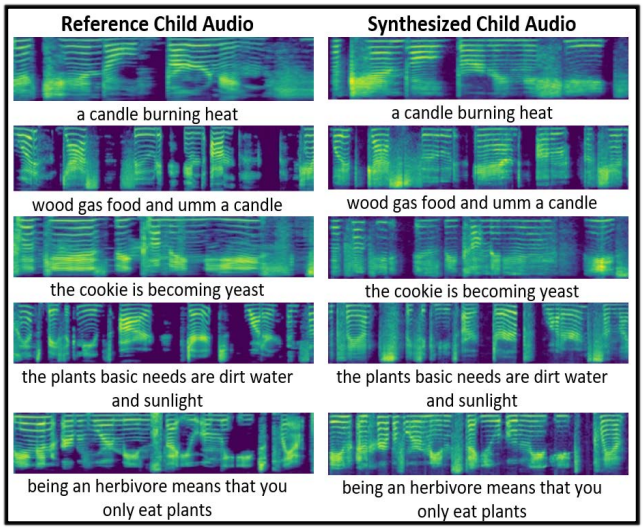


FIGURE 10. Spectrogram comparison between reference and synthesized child audio for 5 audio samples used with MOSNet.

enough data to perform a large-scale evaluation such as a blizzard challenge. This objective evaluation was performed to see the correlation between reference child audio and synthetic child audio. A random set of 50 utterances were selected from the TinyMyST dataset as a part of this inside test. These utterances were used as reference utterances and the corresponding transcripts were used to generate synthetic speech for each of these utterances. This gave us 50 reference and 50 synthetic utterances which were used to calculate MOS using MOSNet. MOS score for 5 samples can be seen in Table 11. The spectrograms for these 5 samples can be seen in Figure 10.

Table 12 shows the overall MOS output for MOSNet. MOS of **2.96** was observed for **reference child audio** and **2.66** for **synthetic child audio**. There is only a 0.3 difference in MOS between reference and synthetic child voices. MOSNet trained on adult speech data is not expected to give MOS ratings correlated with human MOS ratings for child speech data. MOSNet was only used to get a correlation between the

TABLE 11. MOSNet output for 5 samples.

Sample	Reference Child Speech MOS	Synthetic Child Speech MOS
1	2.25	2.80
2	3.08	2.77
3	3.18	2.91
4	3.17	2.51
5	3.17	2.99

TABLE 12. MOSNet output for 50 samples with 95% confidence interval.

Samples	Reference Child Audio MOS	Synthetic Child Audio MOS
50	2.91 ± 0.07	2.60 ± 0.06

reference child audio and synthetic child audio. This gave us a comparison between reference and synthetic child voices as to how close they are to each other in terms of audio features calculated using MOSNet. The results confirmed that MOSNet output for reference child speech and synthetic child speech are very close to each other with a **comparative MOS difference** of **0.3**.

C. SPEAKER SIMILARITY EVALUATION USING A SPEAKER VERIFICATION SYSTEM

Speaker similarity between a synthesized speech and a real speech can be calculated using a Speaker Verification (SV) system. The pretrained speaker encoder from section 2.B.1. was used with a third-party tool[10] to extract and visualize the speaker embeddings. This tool uses cosine distance to calculate the similarity between the two embeddings. The same speakers mentioned in our subjective evaluation (see Table 8) were used for this evaluation. 10 utterances were randomly selected for both real and synthetic speech for each of the 4 speakers mentioned in Table 8. 1 male and 1 female speaker from the LibriSpeech dataset were also added with 10 utterances each to show the speaker similarity comparison between an adult and child speaker. A visualization of this similarity in a 2D projection can be seen in Figure 11, 'gt' is used as a label for the ground truth of the speaker and 'ss' is used as a label for the synthetic speech of the same speaker. 'Adult_Male' and 'Adult_Female' are two randomly selected male and female speakers from the LibriSpeech Dataset.

From Figure 11, it can be inferred that Male, Female, and Child speech have a difference in similarity from each other. Male and Female adult speakers are far apart from each other and from child speakers in this 2D projection of speaker embeddings.

To further comment on the similarity between real child speech and synthetic child speech, the 'child speech' contour from Figure 11 is extended to get a more visual representation of embeddings. This can be seen in Figure 12. The 'gt' labels and very close to 'ss' labels in this 2D projection space.

These embeddings are 256-dimensional feature vectors trained by our speaker encoder. Therefore, cosine similarity was used to further calculate the cross-similarity between

[10]Resemblyzer: https://github.com/resemble-ai/Resemblyzer

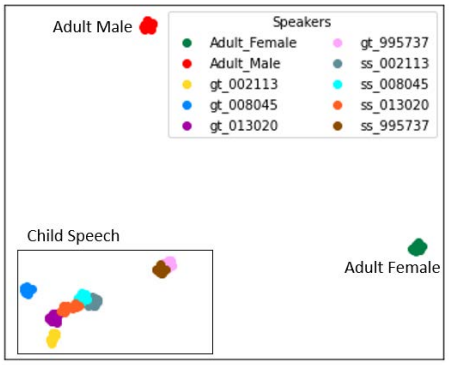


FIGURE 11. Projections of embeddings between different real and synthetic child speech along with adult speech. The child Speech region [both ground truth and synthetic speech] is outlined by a solid black rectangle. The projections include a cluster of 10 voices selected from 10 different speakers. 'ss' refers to synthetic child speech and 'gt' refers to ground truth child speech.

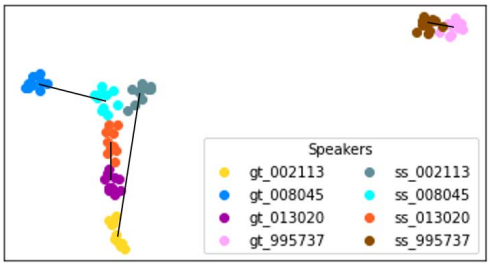


FIGURE 12. Projections of embeddings between different real and synthetic child speech. A solid black line is used to show the distance between the ground truth and synthetic speech from the same speakers. This line was drawn from the centroid of each cluster to show the visual representation of similarity between real and synthetic speech.

each speaker. Each of the 10 Speakers with 10 utterances each (1 Adult Male, 1 Adult Female, 4 Ground Truth Child, and 4 Synthetic Speech Child) were divided into 2 sets A and B. Embeddings are extracted for each of the utterances for each of the sets and averaged together for each speaker. This gave us 10 unique speaker embeddings in sets A and B each for 10 speakers. Cosine similarity is finally used to measure the similarity between each of the 10 speaker embeddings in sets A and B. A plot for the cross similarity between speakers can be seen in Figure 12.

In Figure 13, speaker similarity between synthetic speech and ground truth for speaker '995737' is 0.91, whereas for speakers '013020' and '008045' is approximately around 0.82 and finally for the speaker '002113' is approximately 0.7. This cross-similarity matrix gives us an idea of how close synthetic child voices are in comparison to the real child voice. It also shows us how different an Adult Male and Female Speech is in comparison to a child's speech. Overall, the similarity between most of the child and adult speech is between 0.3-0.4 whereas the similarity between most of the synthetic child speech and ground truth child speech is between a range of 0.65-0.85. Cross-similarity across the diagonal signifies that an utterance in set-A is 95% similar to utterances in set-B having the same speakers. More conclusions can be drawn from Figure 13, however, for the scope of this research, it is only used to show the different

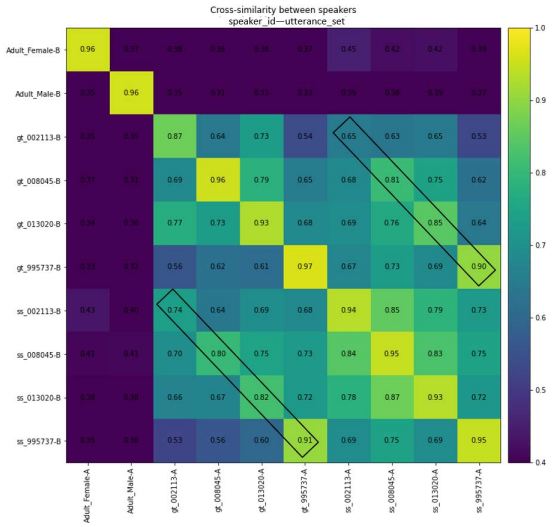


FIGURE 13. Cross-similarity between 10 speakers in Set A and Set B. The rectangular black box represents the similarity between real and synthetic child speech for respective speakers in set-A and set-B. Set-A is along the x-axis and Set-B is along the y-axis. 'ss' represents the synthetic speech and 'gt' represents the ground truth (real) speech.

speaker similarities between real child speech and synthetic child speech and to draw a conclusion that our synthetically generated child speech is very close to real speech in terms of speaker similarity with an average similarity of 81%.

D. OBJECTIVE INTELLIGIBILITY EVALUATION USING A PRETRAINED ASR SYSTEM

A pretrained wav2vec2 model is used to provide verification on synthetic utterances. A comparison of the speech transcription between real and synthetic child voices is presented. Child speech recognition is a challenging task of its own. The ASR on child speech is a part of our future work. Our intent to use this model for this paper is based on the popularity of the model, being SOTA on adult speech. A wav2vec2 model trained on adult speech data is used to provide that comparison. This speech transcription was obtained for the synthetic and real utterances mentioned in Table 8. A random set of 30 utterances for each speaker for both real and synthetic voices are selected. The instruction for using this model is mentioned in their Github.[11]

A comparison of this model is also provided using adult speech by selecting the equal number of adult voices from the LibriSpeech dataset. Word Error Rate (WER) is calculated from the output of wav2vec2 and is mentioned in Table 13. The Flashlight[12] library is used to calculate the WER using Viterbi decoding. No external language model (LM) was used.

From Table 13, it can be inferred that the WER for Adult Speech (Librispeech_test_clean) is 3.43, evidently, due to the model being trained on adult speech data, the WER for real child speech is 15.27 and in comparison, WER for synthetic utterances is 25.63. An ASR model was able to recognize

[11] wav2vec2: https://github.com/pytorch/fairseq
[12] Flashlight: https://github.com/flashlight/flashlight

TABLE 13. WER on adult speech, real child speech and synthetic child speech.

Data type	# of Utterances	WER
Adult Speech [LibriSpeech_test_clean]	120	3.43
Real Child Speech [From MyST]	120	15.27
Synthetic Child Speech [Based on MyST]	120	25.63

75% of the synthetic speech with a relative difference of 10 WER when compared with real child speech recognized by the same model for the same speakers.

V. CONCLUSION AND FUTURE WORK

In this paper, a pipeline for generating synthetic child speech in a limited training data scenario is proposed. A small set of child speech data is created by cleaning an existing child speech dataset and making it suitable for TTS training. A transfer learning approach is used to train our model with adult speech data in a pretraining setting and child speech data as low as 19 hours for fine-tuning. MOSNet based objective evaluation shows a high correlation between real and synthesized child voices. A subjective evaluation method suitable for synthesized child speech is also proposed and demonstrated. Subjective MOS of synthesized voices is observed as 3.95 for speech intelligibility, 3.89 for voice naturalness, and 3.96 for overall voice consistency which is very close to Natural speech MOS. These MOS values tell us about how good the synthesized child voices are. However, voice inconsistency for 'End of phrase quality' containing noise and unintelligible information was also observed. There is scope for improvement for these phrases. WER for synthetic child voices using a pretrained adult speech wav2vec2 ASR model came to be 25.63 as compared to WER of real child voices of 15.27. Synthetic child speech samples can be viewed in our GitHub repository.[13] Multi-speaker TTS can be the key to child speech synthesis with limited training data. Child speakers with speech duration between 5-7 minutes in TTS training gave 81% average cosine similarity with a synthetic speech from the same speakers. This choice of the model allows the TTS to learn useful speaker information which can be leveraged to produce better quality synthetic voices even with limited child speech.

For future work, our aim is to improve this method by incorporating more information to our multi-speaker TTS model such as duration predictor and energy as implemented in FastSpeech2 [12]. The trained vocoder was also finetuned on the TinyMyST dataset. However, there was no significant improvement in the quality of the generated audio waveforms and an additional noise was observed in some of the synthesis. More child speech data would be required to achieve any significant improvement over the quality of the vocoder. It is also intended to implement GAN-based SOTA Vocoders such as HiFi-GAN [19] for future experiments. More experiments such as training a forced aligner using children's voices is also part of our future work. It will help to generate more meaningful alignments for splitting the longer audio files to increase the training dataset. The information collected from our subjective evaluation such as voice consistency in Harvard sentences will be used to improve child speech. This information will be used to collect better TTS-based child speech data based on Harvard sentences to accord with 'end of the phrase' information loss and voice inconsistency observed with our current results. The use of synthetically generated child speech to improve other areas of child speech research such as ASR and speaker recognition will also be investigated in future work. TTS-generated child voices can be used as a data augmentation technique for training these models with additional data. It is also intended to use the subjective evaluation method proposed in this paper for performing all future subjective evaluations with TTS generated child speech.

[13]https://c3imaging.github.io/ChildTTS/

ACKNOWLEDGMENT

The authors would like to thank experts from Xperi-Ireland: Gabriel Costache, George Sterpu, and the rest of the team members for providing their expertise and feedback throughout.

REFERENCES

[1] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T. Y. Liu, "LRSpeech: Extremely low-resource speech synthesis and recognition," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Assoc. Comput. Mach.* New York, NY, USA, 2020, pp. 2802–2812, doi: 10.1145/3394486.
[2] K. R. Prajwal and C V Jawahar, "Data-efficient training strategies for neural TTS systems," in *Proc. 8th ACM IKDD CODS, 26th COMAD*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 223–227, doi: 10.1145/3430984.3431034.
[3] O. Watts, J. Yamagishi, K. Berkling, and S. King, "HMM-based synthesis of child speech," in *Proc. 1st Work. Child, Comput. Interact. (ICMI Post-Conf. Work.*, 2008.
[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009, doi: 10.1016/J.SPECOM.2009.04.004.
[5] P. K. Muthukumar and A. W. Black, "A deep learning approach to data-driven parameterizations for statistical parametric speech synthesis," Carnegie Mellon University Pittsburgh, Pittsburgh, NA, USA, 2014.
[6] O. Watts, J. Yamagishi, S. King, and K. Berkling, "Synthesis of child speech with HMM adaptation and voice conversion," *IEEE Trans. Audio, Speech Language Process.*, vol. 18, no. 5, pp. 1005–1016, Jul. 2010, doi: 10.1109/TASL.2009.2035029.
[7] R. Maia, H. Zen, and M. J. F. Gales, "Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters," in *Proc. SSW*, 2010, pp. 88–93.
[8] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and Q. Le, "Tacotron: Towards end-to-end speech synthesis," 2017, *arxiv:1703.10135*.
[9] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling," 2020, *arxiv:2010.04301*.
[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2756–2761.
[11] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.
[12] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2020, *arXiv:2006.04558*.
[13] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, Jul. 2019, pp. 6706–6713, doi: 10.1609/aaai.v33i01.33016706.

[14] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A non-autoregressive network for text to speech based on flow," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7209–7213, doi: 10.1109/ICASSP40776.2020.9054484.

[15] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," Tech. Rep., 2020.

[16] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[17] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.

[18] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[19] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033.

[20] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2410–2419.

[21] S. Arik, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA, Curran Associates Inc., 2017, pp. 2966–2974.

[22] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4879–4883.

[23] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2016, pp. 5115–5119.

[24] T.-H. Lo, F.-A. Chao, S.-Y. Weng, and B. Chen, "The NTNU system at the interspeech 2020 non-native children's speech ASR challenge," in *Proc. Interspeech*, Oct. 2020, pp. 1–5.

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

[26] E. Cooper, C. I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6184–6188.

[27] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, "MultiSpeech: Multi-speaker text to speech with transformer," in *Proc. Interspeech*, Oct. 2020, pp. 4024–4028.

[28] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6189–6193.

[29] A. Kulkarni, V. Colotte, and D. Jouvet, "Improving latent representation for end-to-end multispeaker expressive text to speech system," Tech. Rep. ffhal-02978485v1f, 2020.

[30] E. Cooper, C.-I. Lai, Y. Yasuda, and J. Yamagishi, "Can speaker augmentation improve multi-speaker end-to-end TTS?" in *Proc. Interspeech*, Oct. 2020, pp. 1–5.

[31] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding," in *Proc. Interspeech*, Sep. 2019, pp. 2105–2109, doi: 10.21437/Interspeech.2019-1632.

[32] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," Univ. California, Berkeley, Tech. Rep., 2017.

[33] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 4480–4490.

[34] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80. Stockholm, Sweden, 2018, pp. 4693–4702.

[35] G. Yeung, R. Fan, and A. Alwan, "Fundamental frequency feature normalization and data augmentation for child speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP2021)*, Toronto, ON, Canada, 2021, pp. 6993–6997, doi: 10.1109/ICASSP39728.2021.9413801.

[36] S. Shahnawazuddin, N. Adiga, H. K. Kathania, and B. T. Sai, "Creating speaker independent ASR system through prosody modification based data augmentation," *Pattern Recognit. Lett.*, vol. 131, pp. 213–218, Mar. 2020, doi: 10.1016/j.patrec.2019.12.019.

[37] S. Shahnawazuddin, R. Sinha, and G. Pradhan, "Pitch-normalized acoustic features for robust children's speech recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1128–1132, Aug. 2017, doi: 10.1109/LSP.2017.2705085.

[38] S. Lee, A. Potamianos, and S. S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. Eurospeech*, 1997.

[39] S. Shahnawazuddin, N. Adiga, and H. K. Kathania, "Effect of prosody modification on Children's ASR," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1749–1753, Nov. 2017, doi: 10.1109/LSP.2017.2756347.

[40] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. 2nd Workshop Child, Comput. Interact. (WOCCI)*, 2009, pp. 1–8.

[41] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic normalization of children's speech," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003.

[42] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2006, pp. I–I, doi: 10.1109/ICASSP.2006.1660040.

[43] C. Li, Y. Qian, and M. Key, "Prosody usage optimization for children speech recognition with zero resource children speech," in *Proc. Interspeech*, 2019, pp. 3446–3450.

[44] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468.

[45] W. Ward, R. Cole, and S. Pradhan, "My science tutor and the MyST corpus," Tech. Rep., 2019.

[46] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, Aug. 2017, pp. 1–5.

[47] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.

[48] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Centre Speech Technol. Res. (CSTR), Univ. Edinburgh, Tech. Rep., 2019, doi: 10.7488/ds/2645.

[49] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," 2019, *arXiv:1904.02882*.

[50] *The LJ Speech Dataset*. Accessed: Mar. 15, 2021. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[51] *CorentinJ/Real-Time-Voice-Cloning: Clone a Voice in 5 Seconds to Generate Arbitrary Speech in Real-Time*. Accessed: May 27, 2021. [Online]. Available: https://github.com/CorentinJ/Real-Time-Voice-Cloning

[52] L. Mcinnes and J. Healy, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[53] *Fatchord/WaveRNN: WaveRNN Vocoder + TTS*. Accessed: May 27, 2021. [Online]. Available: https://github.com/fatchord/WaveRNN

[54] P.-C. Hsu, C.-H. Wang, A. T. Liu, and H.-Y. Lee, "Towards robust neural vocoding for speech generation: A survey," 2019, *arXiv:1912.02461*.

[55] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards achieving robust universal neural vocoding," 2018, *arXiv:1811.06292*.

[56] P. L. Tobing and T. Toda, "High-fidelity and low-latency universal neural vocoder based on multiband WaveRNN with data-driven linear prediction for discrete waveform modeling," in *Proc. Interspeech*, 2021, p. 2105.

[57] D. Paul, Y. Pantazis, and Y. Stylianou, "Speaker conditional WaveRNN: Towards universal neural vocoder for unseen speaker and recording conditions," in *Proc. Interspeech*, 2020.

[58] Z. Cai, C. Zhang, and M. Li, "From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint," in *Proc. Interspeech*, 2020, pp. 3974–3978, doi: 10.21437/Interspeech.2020-1032.

[59] B. Naderi and R. Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," in *Proc. Interspeech*, 2020, pp. 2862–2866, doi: 10.21437/Interspeech.2020-2665.
[60] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoustics*, vol. AU-17, no. 3, pp. 225–246, Jun. 1969.
[61] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale," *Comput. Speech Lang.*, vol. 19, no. 1, pp. 55–83, Jan. 2005, doi: 10.1016/J.CSL.2003.12.001.
[62] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," Tech. Rep., Oct. 2021.
[63] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4214–4217, doi: 10.1109/ICASSP.2010.5495701.
[64] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010, doi: 10.1109/TASL.2010.2052247.
[65] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 749–752, doi: 10.1109/ICASSP.2001.941023.
[66] C.-C. Lo, S. W. Fu, W. C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H. M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545, doi: 10.21437/Interspeech.2019-2003.
[67] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "WebMUSHRA—A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, p. 8, Feb. 2018, doi: 10.5334/jors.187.
[68] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 2416–2419.
[69] S. Zielinski, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential biases in MUSHRA listening tests," in *Proc. Audio Eng. Soc. 123rd Audio Eng. Soc. Conv.*, vol. 2, 2007, pp. 1–10.
[70] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations," in *Proc. Interspeech*, 2015, pp, 3476–3480, doi: 10.21437/Interspeech.2015-689.

RISHABH JAIN (Graduate Student Member, IEEE) received the B.Tech. degree in computer science and engineering from the Vellore Institute of Technology (VIT), in 2019, and the M.S. degree in data analytics from the National University of Ireland Galway (NUIG), in 2020, where he is currently pursuing the Ph.D. degree. He is also working as a Research Assistant at NUIG under DAVID (Data-Center Audio/Visual Intelligence on-Device) Project. His research interests include machine learning and artificial intelligence specifically in domain of speech understanding, text-to-speech, speaker recognition, and automatic speech recognition.

MARIAM YAHAYAH YIWERE received the Bachelor of Science degree from the Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, in 2012, and the Master of Engineering and Ph.D. degrees from the Department of Computer Engineering, Hanbat National University, South Korea, in August 2015 and February 2020, respectively. Since October 2020, she has been working on the DTIF/DAVID Project as a Postdoctoral Researcher with the College of Science and Engineering, National University of Ireland Galway, Galway. Her research interests include text-to-speech synthesis, speaker recognition and verification, sound source localization, deep learning, and computer vision.

DAN BIGIOI (Graduate Student Member, IEEE) received the bachelor's degree in electronic and computer engineering from the National University of Ireland Galway, in 2020. Upon graduating, he worked as a Research Assistant at NUIG studying the text-to-speech and speaker recognition methods under the DAVID (Data-Center Audio/Visual Intelligence on-Device) Project. Currently, he is working on his Ph.D. at NUIG, sponsored by D-REAL and the SFI Centre for Research Training in Digitally Enhanced Reality. His research interests include novel deep learning-based techniques for automatic speech dubbing and discovering new ways to process multimodal audio/visual data.

PETER CORCORAN (Fellow, IEEE) is currently holding the Personal Chair of Electronic Engineering with the College of Science and Engineering, National University of Ireland Galway (NUIG). He was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He has more than 600 cited technical publications and patents, more than 120 peer-reviewed journal articles, 160 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is also a member of the IEEE Consumer Technology Society for more than 25 years and the Founding Editor of *IEEE Consumer Electronics Magazine*.

HORIA CUCU (Member, IEEE) received the B.S. and M.S. degrees in applied electronics and the Ph.D. degree in electronics and telecommunication engineering from the University Politehnica of Bucharest (UPB), Romania, in 2008 and 2011, respectively.

From 2010 to 2017, he was a Teaching Assistant and then a Lecturer at UPB, where he is currently working as an Associate Professor. In this position, he authored over 75 scientific papers in international conferences and journals, served as the project director for seven research projects, and contributed as a researcher to ten other research grants. He holds two patents. In addition, he founded and leads Zevo Technology, a speech start-up dedicated to integrating state-of-the-art speech technologies in various commercial applications. His research interests include machine/deep learning and artificial intelligence, with a special focus on automatic speech and speaker recognition, text-to-speech synthesis, and speech emotion recognition.

Dr. Cucu was awarded the Romanian Academy Prize "Mihail Drăgănescu" (2016) for Outstanding Research Contributions in Spoken Language Technology, after developing the first large-vocabulary automatic speech recognition system for the Romanian language.

Appendix B

A WAV2VEC2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition.

Authors: Rishabh Jain (RJ), Andrei Barcovschi (AB), Mariam Yiwere (MY), Dan Bigioi (DP), Peter Corcoran (PC) and Horia Cucu (HC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	RJ: 80%, PC:20%
Experiments and Implementation	RJ: 80%, AB:10%, MY:10%
Background	RJ: 90%, AB:10%
Manuscript Preparation	RJ: 70%, AB: 5%, MY: 5%, PC: 10%, HC: 10%

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 April 2023, accepted 7 May 2023, date of publication 10 May 2023, date of current version 17 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3275106

RESEARCH ARTICLE

A WAV2VEC2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition

RISHABH JAIN[1], (Graduate Student Member, IEEE), ANDREI BARCOVSCHI[1], MARIAM YAHAYAH YIWERE[1], DAN BIGIOI[1], (Graduate Student Member, IEEE), PETER CORCORAN[1], (Fellow, IEEE), AND HORIA CUCU[2], (Member, IEEE)

[1]School of Electrical and Electronics Engineering, University of Galway, Galway, H91 TK33 Ireland
[2]Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, 060042 Bucuresti, Romania

Corresponding author: Rishabh Jain (rishabh.jain@universityofgalway.ie)

This work was supported in part by the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project (2020–2023) funded by the Disruptive Technologies Innovation Fund (DTIF), College of Science and Engineering Ph.D. Research Scholarship, University of Galway, Science Foundation Ireland (SFI) Center for Research Training in Digitally Enhanced Reality under Grant 18/CRT/6224; and in part by SFI ADAPT Center for Digital Media Research under Grant 13/RC/2106_P2.

ABSTRACT Despite recent advancements in deep learning technologies, Child Speech Recognition remains a challenging task. Current Automatic Speech Recognition (ASR) models require substantial amounts of annotated data for training, which is scarce. In this work, we explore using the ASR model, wav2vec2, with different pretraining and finetuning configurations for self-supervised learning (SSL) toward improving automatic child speech recognition. The pretrained wav2vec2 models were finetuned using different amounts of child speech training data, adult speech data, and a combination of both, to discover the optimum amount of data required to finetune the model for the task of child ASR. Our trained model achieves the best Word Error Rate (WER) of 7.42 on the MyST child speech dataset, 2.91 on the PFSTAR dataset and 12.77 on the CMU KIDS dataset using cleaned variants of each dataset. Our models outperformed the unmodified wav2vec2 BASE 960 on child speech using as little as 10 hours of child speech data in finetuning. The analysis of different types of training data and their effect on inference is provided by using a combination of custom datasets in pretraining, finetuning and inference. These 'cleaned' datasets are provided for use by other researchers to provide comparisons with our results.

INDEX TERMS Child speech recognition, self-supervised learning, wav2vec2, automatic speech recognition, MyST dataset, PFSTAR dataset, CMU_kids dataset.

I. INTRODUCTION

Current deep learning-based automatic speech recognition (ASR) models perform remarkably well on adult speech data. However, they struggle when it comes to recognizing speech from children. Models such as wav2vec2, Deep Speech 2, ContextNet, and others [1], [2], [3], [4], [5], [6], [7] all achieve impressive results on adult speech datasets such as LibriSpeech (∼1000h), TIMIT (5.4h), LJSpeech (∼24h), MediaSpeech (∼10h), and more. This is due in no small part to the vast amounts of annotated adult speech data available for training such models and the ease with which it can be obtained. However, when it comes to child speech recognition, State-Of-The-Art (SOTA) ASR models trained on adult data perform quite poorly on child voice datasets. This is due to the inherent differences between adult and children's voices. A child's voice is quite different from an adult's voice [8], [9] in terms of pitch, linguistic and acoustic features, ability to understand and pronounce words, high fundamental frequency, and shorter vocal tract length.

In addition, it is a challenging task to collect and annotate child speech data in comparison to adult speech data which

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

can be acquired from various sources such as movies, news broadcasts, audiobooks, internet, etc. Even if child speech can be collected from such sources, providing accurate annotations remains challenging. When compared to adult voice datasets, child voice datasets are quite limited [10].

ASR is an important and useful tool for speech researchers. It forms the basis of speech understanding [11] when combined with advanced language models, but also finds applications in generative models and for training improved Text-To-Speech (TTS) models [12], [13], [14]. The interrelationship between ASR and TTS is further described in [15]. As our underlying motivation is related to TTS models and their finetuning, we cleaned the publicly available datasets used in this research to provide improved annotations for TTS models.

A. RELATED WORKS

In the past few years, there have been many different approaches to improving the performance of automatic child speech recognition systems [16]. Most of these approaches consist of various data augmentation techniques for increasing the amount of usable training data. Text-to-Speech based data augmentations as introduced by [14] and [17], where ASR models are finetuned using synthetic data, have not shown significant increases in the accuracy of child ASR. Generative Adversarial Network (GAN) based augmentation [18], [19], [20] has also been explored to increase the amount of labeled data with acoustic attributes like those of child speech. Some of the other popular augmentation approaches include Vocal Tract Length Perturbation [21], Fundamental frequency feature normalization [22], out-of-domain data augmentation using Stochastic Feature Mapping (SFM) [23], and data processing-based augmentations [24] such as Speed Perturbation, Pitch Perturbation, Tempo Perturbation, Volume Perturbation, Reverberation Perturbation, and Spectral Perturbation. Spectrogram Augmentation also seems promising for improving the performance of ASR systems [25], [26]. Each of these methods shows improvements in child ASR accuracy, however, they still require corresponding labeled annotations to speech data.

Another recent trend is the use of transfer learning approaches for improving the recognition in child ASR for features adaptability from adult to child speech. The authors in [27] perform extensive analysis to understand the effect of the amount of adaptation data, different Deep Neural Network (DNN) transfer learning configurations, and their impact on different age groups for improving child ASR. In [28], the authors explored the use of a two-step training strategy, which involves multilingual pretraining followed by transfer learning, for improving the performance of ASR systems on child speech.

Each of these methods show some improvements in child ASR accuracy, however, they still require corresponding labeled annotations to speech. A recent review of child ASRs [21] determined that most of these SOTA methods are supervised learning approaches. The authors in [29] show the performance of various supervised learning approaches for ASR in child speech. They compared the performance of end-to-end ASR systems with that of Deep Neural Network-Hidden Markov Model (DNN-HMM) hybrid systems. Another paper [30] studied the performance of Factored Time Delay Neural Networks (TDNN-F) with traditional and SOTA systems for ASR of child speech. These supervised approaches rely on labeled child speech data during training for the task of ASR.

As there is a distinct lack of labeled child speech data compared to adult, approaches that utilize unsupervised [31] and self-supervised learning [1] were explored for this paper. Therefore, the goal of this work is to present a method to incorporate unlabeled child speech data into the training procedure of a typical ASR model while also making use of abundant, labelled, and unlabeled adult speech data to improve the overall accuracy of ASR models on child speech.

B. SELF-SUPERVISED LEARNING FOR CHILD ASR

Self-supervised learning (SSL) has emerged as a paradigm to learn general data representations from substantial amounts of unlabeled examples allowing one to then fine-tune models on small amounts of labeled data. The use of SSL for child ASR was first seen at Interspeech2021, where a model using SSL [32] received first place for non-native child speech challenge. A similar use case [24] was also presented in the SLT 2021 children speech recognition challenge [33]. Another approach is used in [34], where the author uses a bidirectional unsupervised model pretraining with child speech ASR. After reviewing various approaches to SSL, wav2vec2 [1] was chosen for this paper. Wav2vec2 shows that using SSL for the task of ASR provides improvements over SOTA supervised learning approaches.

At the time of working on this paper, many applications of the wav2vec2 model for child ASR were observed. The authors in [35] propose the use of a transformer model pretrained on adult speech to achieve SOTA results on children's dataset. Reference [36] a comparison between different SSL approaches for child speech recognition tasks. In [37], authors proposed a Domain Responsible Adaptation and Fine-Tuning (DRAFT) framework to address the domain shift between adult speech used for pretraining and child speech used for finetuning. They use wav2vec2 along with other SSL methods to examine the cross-domain transfer between different children's datasets.

This paper explores various pretraining and finetuning configurations with different combinations of adult and child speech datasets using wav2vec2 speech representations. Three child speech datasets were used in this study. These datasets were cleaned and preprocessed to make them usable for ASR. We also report the best results on different child speech validation. The ideal data requirement for pretraining and finetuning in a low-data scenario was also explored in this paper by observing the relation/pattern of performance

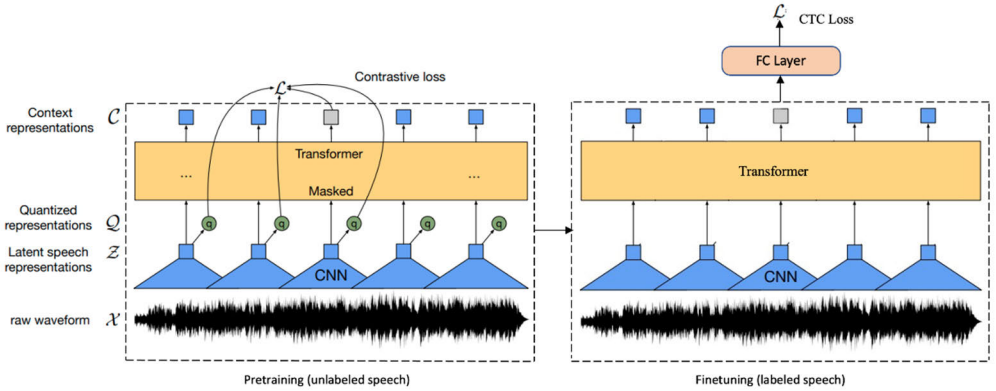


FIGURE 1. Pretraining and finetuning steps in Wav2vec2 (from [1]).

in different datasets used. The rest of this paper is organized as follows: Section II describes the model architecture. Section III introduces the datasets used for this paper. Section IV includes the codebase and experiments. Section V describes the results. Conclusions are presented in Section VI.

II. TRAINING METHODOLOGY FOR SSL

The wav2vec2 model [1] is used to extract speech representations from raw audio files in a self-supervised learning scenario and use these representations for ASR-specific tasks. Wav2vec2 is used in this paper as it can achieve SOTA results when trained on a large amount of unlabeled speech data and finetuned on labeled data as small as 10 minutes. This is ideal for our task, as it is much easier to obtain significant amounts of unlabeled child speech data than gather accurately labeled data.

As it is a two-step training method (See Figure 1), the first step includes a pretraining step in which the model is trained with a large amount of unlabeled data. The second step includes finetuning on labeled data using Connectionist Temporal Classification (CTC) loss [38] for downstream ASR tasks. As the model learns SSL speech representation in pretraining, it can be trained using large quantities of unlabeled speech data and can be finetuned with only a small amount of labeled data. This way, the problem of scarcity of child speech is solved as we can train the 'pretraining' model with a combination of unlabeled speech data and it can also be used to learn speech representations from adult speech datasets making use of the abundant adult speech data.

A. PRETRAINING

The pretraining stage of the wav2vec2 model consists of a feature encoder, context network, and quantization module. The CNN feature extractor takes the raw audio waveform as input and passes it through a series of 1D convolutional layers to extract high-level representations from the waveform. The output of the feature extractor is a sequence of feature vectors that represent the input waveform. The context network is a transformer-based encoder which takes this sequence of feature vectors and processes them using a stack of transformer layers. The transformer layers in wav2vec 2.0 use a self-attention mechanism allowing the model to capture long-range dependencies in the input data. The quantization module consists of a codebook of fixed vectors, where each input feature vector is assigned to the closest codebook vector. Gumbel softmax function [39] is used to choose the quantized representation from multiple codebooks. After quantization, the discrete symbols are passed through a transformer encoder, which learns to encode the sequence of symbols into a fixed-length representation that can be used for downstream tasks such as speech recognition. Since the process involves mapping continuous values to discrete values, it makes the model to be more efficient for training and inference.

The contrastive loss function in Wav2vec2 is applied after the quantization is performed. It is used to train the model to produce embeddings that capture useful features of speech signals. This is followed by a diversity loss which encourages similar feature vectors to be closer together and dissimilar feature vectors to be farther apart. By minimizing these losses, Wav2vec2 can learn to produce embeddings that are effective for downstream speech recognition tasks.

Experiments' configurations are provided as the BASE and LARGE models. The configurations differ in transformer block size but use the same size for the encoder. The feature encoder contains seven blocks with each block having strides of (5,2,2,2,2,2,2) and kernel widths of (10,3,3,3,3,2,2) and output temporal convolution of 512 channels. The context network of the BASE model contains 12 transformer blocks, each block with a 512-dim model, 8 attention heads, and a 2048-dim feed-forward inner layer, while the LARGE model contains 24 transformer blocks with model dimensions 1024, inner dimensions 4096, and 16 attention heads. We use 4 NVIDIA Tesla V100 GPUs to pretrain the model. Model pretraining was optimized using ADAM [40]. During the first 8% of updates, the learning rate warms up to a peak of 5×10^{-4} for BASE and 3×10^{-4} for LARGE, and then it linearly decays. We use both BASE and LARGE models

according to dataset size used for pretraining. BASE models contain 93M parameters and LARGE models contain 317M parameters.

B. FINETUNING

For finetuning, 29 target letters were used (from the Librispeech dataset) as provided by the authors in wav2vec2 [1]. Models are optimized by minimizing CTC loss [38] for ASR task. A modified version of SpecAugment [25] is applied as masking to timestamps and channels to reduce the overfitting and improve the recognition robustness. We fine-tune on one V100 GPU. For the first 1000 updates, only the final output classifier was trained, after which the Transformer block was also trained. The feature encoder was frozen during finetuning training. We also use different finetuning configurations depending on the size of finetuning datasets. The hyperparameters are kept the same as provided by the wav2vec2 authors [1]. The learning rate changes according to the dataset size as documented by the authors of wav2vec2 [1].

As the goal of this study is to evaluate the performance of self-supervised speech representations, it was decided not to incorporate a language model in this research. Additionally, previous research has shown that the best results for children's ASR systems were achieved without the use of an external language model [29]. Language model adaptation for child speech is also an unexplored research area. Child speech would require a specialized trained language model for best results. As there isn't any definitive publicly available language model for child speech, we consider this as a part of the future research topic.

III. DATASET DESCRIPTION AND USAGE

The datasets are divided according to their usage. The child speech data used in this paper include MyST Corpus [41], CMU_Kids [42] and PF-STAR [43]. Adult Speech datasets include Librilight [44], LibriTTS [45], and LibriSpeech [46].

A. DATASET DESCRIPTION

Below we provide a description of the datasets used in this paper:

1) LIBRISPEECH [46]

Librispeech is an adult speech dataset with approximately 1000 hours of recorded audio with a sampling rate of 16Khz. The data is derived from read audiobooks from the LibriVox project. The data is carefully segmented, aligned, and used popularly in speech research.

2) LIBRILIGHT [44]

Librilight is an adult speech dataset used as a benchmark for training speech recognition systems with limited or no supervision. It contains 60,000 hours of unlabeled adult speech extracted from audiobooks. It was mentioned in the wav2vec2 paper [1] and used by the authors.

3) LibriTTS [45]

The LibriTTS dataset is a large-scale dataset for training TTS models and is a subset of the Librispeech dataset. It consists of approximately 560 hours of high-quality audio and text transcriptions from audiobooks. This dataset is used here for inference over adult speech as it is a clean and noise-free dataset. The 'dev-clean' segment of the LibriTTS dataset which contains over 8.9 hours of clean adult speech. It is also widely used as a baseline in the validation of ASR and TTS experiments.

4) MY SCIENCE TUTOR (MySt) CHILD SPEECH [41]

The MyST (My Science Tutor) Children's Speech Corpus consists of 393 hours of American English children's speech with a total of 228,874 utterances. The speech was collected from 1371 third, fourth and fifth-grade students. 45% of the utterances have been transcribed at the word level amounting to 197 hours. This dataset is used in this paper as it's the largest open-source corpus of child speech available for research use.

5) PF-STAR CORPUS OF BRITISH ENGLISH CHILD SPEECH [43]

This corpus contains British English child speech from 158 children aged 4 to 14 years. The recordings are divided into a training set (7.5 hours), an evaluation set (1 hour) and a test set (5.6 hours). The corpus was collected at three locations: a university laboratory and two primary schools. It contains both read and spontaneous child speech with transcriptions.

6) CMU KIDS [42]

CMU KIDS Corpus contains read-aloud sentences by children. It was created to provide training data for the SPHINX II automatic speech recognizer at Carnegie Mellon University. It contains 9 hours of American English child speech. The dataset contains 24 male and 52 female speakers having a total of 5180 utterances.

B. DATASET CLEANING AND PROCESSING

All speech data was converted into a 16-bit mono channel with a 16Khz sampling rate, wherever required. All the transcriptions were cleaned and normalized to remove abbreviations, punctuations, whitespaces, etc. and all the characters were changed to uppercase. All the non-linguistic annotation symbols (in child speech datasets) such as "<unk>, sil, hmm, <breath>, <noise>, <indiscernible>, [ze-], [cham-], [***ision], etc." were removed and only alphanumeric characters were retained in the transcript. This was done for all the labeled data used in this paper. Child datasets required further cleaning and pre-processing as follows:

1) MYST CLEANUP

We use the transcribed portion of MyST dataset containing over 197 hours of speech data presented in .trn file

format. The MyST dataset contained a lot of noisy and non-meaningful sentences such as:

- <silence> I'm i don't know <noise> actually
- <whisper> sending go back (*)
- <whisper> what's this one <side_speech> it's an
- give me that <indiscernible> a circuit is a pathway
- <laugh> yeah yeah

The content between '<' and '>' tags were removed from all the transcriptions along with the tags themselves. All the cleaned text files were saved in a .txt format. On further inspection, it was observed that samples below 10 seconds in length generally contained non-meaningful, noisy speech, and data above 20 seconds would lead to GPU running out of memory. Therefore, 10-20 seconds long speech samples from transcribed MyST were selected for finetuning. A final cleaning was performed by manually removing some of the non-meaningful utterances by listening to audio files and going through the transcripts, which amounted to a total of 65 hours of clean data. The data was then randomly split into two groups having 55 hours of data for training and 10 hours for testing as can be seen in Table 1.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="(unknown)" audio_filename="digits1" version="2" version_date="031105">
<Episode>
<Section type="report" startTime="0" endTime="36.497">
<Turn startTime="0" endTime="36.497">
<Sync time="0"/>
sil
<Sync time="0.985"/>
five
<Sync time="1.735"/>
two
<Sync time="2.289"/>
four
<Sync time="2.852"/>
sp
<Sync time="3.445"/>
seven
<Sync time="4.098"/>
sp
<Sync time="4.258"/>
five
<Sync time="4.727"/>
nine
<Sync time="5.289"/>
sp
<Sync time="5.883"/>
one
<Sync time="6.414"/>
oh
<Event desc="error_wrong_word(s)" type="noise" extent="instantaneous"/>
<Sync time="6.914"/>
one
<Sync time="7.289"/>
sil
```

FIGURE 2. Example of '.trs' file in the pfstar dataset. The content in this image was segmented into 'five two four', 'seven', 'five nine', and 'one oh one'. The image is provided to show an example of how transcripts data were made available using '.trs' transcriber old format.

2) PFSTAR CLEANUP

The PFSTAR corpus also contained a lot of non-meaningful utterances and noisy data samples. The dataset comes with '.trs' transcription files, containing time-aligned text information (see Figure 2). These timestamps were used to further segment the data into small audio chunks and remove noise from the dataset. The 'sp' tag from the transcription was used to divide the long transcripts into smaller segments. The corresponding time information was used to segment the long audio files into smaller chunks using FFmpeg[1] and Python. The audio files from the PFSTAR dataset which were 30-70 seconds long were segmented into smaller audio chunks of 5-20 seconds in duration. This segmentation led to 12 hours of clean, usable PFSTAR data, which was further divided into 2 sets: PFS_10h with 10 hours of data (for training) and PFS_test with 2 hours of data (for inference). The final audio data was saved in .wav format and transcriptions in .txt format.

[1]FFmpeg: https://ffmpeg.org/

3) CMUKIDS CLEANUP

CMU_Kids dataset also contains a lot of noisy and incomprehensible child speech. The transcriptions are provided in a '.trn' file format and audio files in a '.sph' format. The data was cleaned in a similar way to MyST by removing all the unrequired tags and non-textual information from the transcripts. For example, ''they [begin_noise] kept a few [end_noise] butterflies in [noise]'' was converted to ''they kept a few butterflies in''. A few more examples can be seen below:

- [begin_noise] cages [end_noise] to lay more eggs [noise] [sil]
  -> cages to lay more eggs
- a [begin_noise] blue butterfly [end_noise] /F L R UW/ [human_noise] flew by [human_noise] [human_noise]
  -> a blue butterfly flew by

The cleaned dataset contained all the audio files in '.wav' format and all transcribed speech in '.txt' format as needed for our training. The total amount of CMU_Kids dataset amounted to 9 hours which was used during inference only.

C. DATASET USAGE

The dataset usage is mentioned in Table 1. The 'Usage' column indicates whether the dataset was used for

TABLE 1. Dataset description for pretraining, finetuning and inference.

Usage	Dataset	Duration	Type
Pretraining [Unlabeled data]	MyST_complete	393 hrs	Child
	Librispeech	960 hrs	Adult
	Libri-light	60k hrs	Adult
Finetuning [Labeled data]	MyST_10m	10 mins	Child
	MyST_1h	1 hr	Child
	MyST_10h	10 hrs	Child
	MyST_55h	55 hrs	Child
	PFS_10m	10 mins	Child
	PFS_1h	1 hr	Child
	PFS_10h	10 hrs	Child
	LS_10m	10 mins	Adult
	LS_100h	100 hrs	Adult
	LS_960h	960 hrs	Adult
Inference [Labeled data]	MyST_test	10 hrs	Child
	PFS_test	2 hrs	Child
	CMU_Kids	9 hrs	Child
	LibriTTS 'dev-clean'	8.9 hrs	Adult

TABLE 2. Group-A: WER for different pretraining (Adult speech datasets) and finetuning (Adult speech dataset) experiments on the MYST, PF-STAR, CMU KIDS and LIBRITTS 'dev-clean' datasets.

Group	Model ID	Pretraining Model Configuration	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMU_KIDS	WER dev_clean
GROUP - A	1	BASE	Librispeech	LS_10m	31.48	30.05	33.38	15.90
	2			LS_100h	17.82	15.96	18.73	4.16
	3			LS_960h	**15.41**	**11.20**	**16.33**	**3.40**
	-	*Average (Group – A, BASE)*			*21.57*	*19.07*	*22.81*	*7.82*
	4	LARGE	Librilight	LS_10m	26.47	27.14	29.37	15.35
	5			LS_100h	13.15	11.63	16.18	3.79
	6			LS_960h	**12.50**	**8.56**	**14.85**	**3.28**
	-	*Average (Group – A, LARGE)*			17.37	15.78	20.13	7.47

pretraining, finetuning, or inference. The 'Type' column specifies whether the dataset consists of child or adult speech. Dataset name is mentioned in 'Dataset' column while amount (in hours/minutes) is mentioned under 'Duration' column.

Pretraining datasets only consists of audio files and doesn't require any transcript/labelled data during training. Finetuning data consists of audio files along with labelled transcripts. The size of the finetuning datasets was chosen as instructed in wav2vec2 [1], and to keep it consistent with their methodology. A similar distribution was maintained for finetuning with child speech datasets (wherever possible). The data was segmented randomly for creating various finetuning subsets.

IV. CODEBASE AND EXPERIMENTS

A. CODEBASE AND HYPERPARAMETERS

The wav2vec2 implementation provided by the fairseq[2] framework is used for our experiments. Hyperparameters were kept the same for both BASE and LARGE pretraining configurations as provided by the wav2vec2 authors. Finetuning configurations were also kept consistent with the finetuning dataset size used. Data cleaning and data processing scripts were created using FFmpeg and Python-based tools such as pydub and scipy. All the training checkpoints are made available on our GitHub page[3] and can be used directly with the model implementation from fairseq. See note[4] for more information on data cleaning scripts and dataset availability.

B. EXPERIMENTS

Experiments were divided into five groups, Group-A, B, C, D and E. ASR performance is measured in terms of Word Error Rate (WER) on different adult and child speech datasets. Child speech datasets used in inference include unseen MyST_test, PFS_test and CMU_Kids, and adult speech dataset include LibriTTS 'dev-clean'. These datasets are common for all groups during inference tests. All the groups of experiments (except Group-C) use two model configurations, namely BASE and LARGE. The BASE configuration includes 960 hours of Librispeech pretraining data and the LARGE configuration includes 60k hours of Librilight data, which is 60 times as much pretraining data as in the BASE configuration. This enables an assessment of the importance of the original training data size for the wav2vec2 model.

For Group-A (Table 2), the finetuned checkpoints provided by the wav2vec2 repository were used for inference. Each of the BASE and LARGE configurations were finetuned with 10 minutes, 100 hours, and 960 hours of Librispeech. For Group-B (Table 3), the pretrained model is finetuned with 10 minutes, 1 hour, 10 hours, and 55 hours of MyST child speech data. In Group-C (Table 3), the Librispeech and MyST datasets having 960 hours of adult speech and 393 hours of child speech data, respectively, are used for pretraining. The model is then finetuned over different amounts of the MyST dataset (similar to Group-B). We only use BASE configuration for this experiment. Group-D (Table 4) uses PFSTAR dataset for finetuning instead of the MyST dataset, and both BASE and LARGE configuration are finetuned with 10 minutes, 1 hour and 10 hours of PFSTAR child speech dataset. Group-E (Table 5) uses a mix of different datasets in the finetuning. A mix of the MyST_55h, PFS_10h, and LS_960 datasets was used. Finetuning mix included LS_960h+MyST_55h, LS_960h+PFS_10h, MyST_55h+PFS_10h and LS_960h+MyST_55h+ PFS_10h. These experiments were performed to see the cross-domain correlation in WER across different finetuning datasets.

Note that we did not train any models from scratch with child speech data alone as there is not sufficient publicly available child speech data to learn any meaningful speech representations from child speech alone. This is discussed in more detail in section V.

V. RESULTS AND DISCUSSION

A. MAIN RESULTS FROM THE GROUP EXPERIMENTS

Results from group experiments are presented in Tables – 2, 3, 4, and 5, with lowest WERs highlighted in bold.

[2]https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

[3]https://github.com/C3Imaging/childASR_w2v2

[4]**Note:** We only make the basic data cleaning scripts available in the GitHub. Researchers trying to replicate our work can email us and get access to other research material. For access to respectively cleaner versions of datasets used in this paper, researchers can buy their own license for the original datasets (where required), and on providing proof of that license, can get access to our 'clean' versions.

TABLE 3. Group-B and Group-C: WER for different pretraining (adult and child speech datasets) and finetuning (MyST child speech dataset) combinations on the MYST, PF-STAR, CMU KIDS and LIBRITTS 'dev-clean' datasets.

Group	Model ID	Pretraining Model Configuration	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMU_KIDS	WER dev_clean
GROUP-B	7	BASE	Librispeech	MyST_10m	28.84	41.34	34.18	21.45
	8			MyST_1h	18.75	31.84	23.13	13.91
	9			MyST_10h	13.46	28.68	19.59	10.94
	10			MyST_55h	**8.13**	**14.77**	**16.47**	**7.72**
	-	*Average (Group – B, BASE)*			*17.29*	*29.16*	*23.34*	*13.51*
	11	LARGE	Librilight	MyST_10m	33.01	44.36	39.91	46.45
	12			MyST_1h	14.91	26.21	18.74	11.59
	13			MyST_10h	12.92	25.05	17.72	10.04
	14			MyST_55h	**7.51**	**12.46**	**15.25**	**6.43**
	-	*Average (Group – B, LARGE)*			*17.08*	*27.02*	*22.91*	*18.62*
GROUP-C	15	BASE	Librispeech MyST_Complete	MyST_10m	29.16	45.71	37.56	35.39
	16			MyST_1h	21.89	38.53	29.03	20.45
	17			MyST_10h	16.18	32.95	25.06	16.83
	18			MyST_55h	**10.34**	**25.47**	**23.15**	**13.48**
	-	*Average (Group – C, BASE)*			*19.39*	*35.67*	*28.7*	*21.53*

TABLE 4. Group-D: WER for different pretraining (adult speech datasets) and finetuning (PFstar child speech dataset) combinations on the MYST, PF-STAR, CMU KIDS and LIBRITTS 'dev-clean' datasets.

Group	Model ID	Pretraining Model Configuration	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMU_KIDS	WER dev_clean
GROUP-D	19	BASE	Librispeech	PFS_10m	35.91	16.43	33.53	30.43
	20			PFS_1h	33.52	7.36	29.55	16.61
	21			PFS_10h	**31.86**	**3.48**	**27.49**	**13.95**
	-	*Average (Group – D, BASE)*			*33.76*	*9.09*	*30.19*	*20.33*
	22	LARGE	Librilight	PFS_10m	37.10	16.78	35.13	23.85
	23			PFS_1h	30.81	14.19	28.54	21.89
	24			PFS_10h	**27.17**	**3.50**	**21.35**	**11.60**
	-	*Average (Group – D, LARGE)*			31.69	11.49	28.34	19.11

TABLE 5. Group-E: WER for different pretraining (adult datasets) and finetuning (adult and child speech datasets) combinations on the MYST, PF-STAR, CMU KIDS and LIBRITTS 'dev-clean' datasets.

Group	Model ID	Pretraining Model Configuration	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMUKIDS	WER dev_clean
GROUP-E	25	BASE	Librispeech	LS_960h, MyST_55h	8.18	12.17	14.12	1.24
	26			LS_960h, PFS_10h	15.42	3.74	15.31	1.41
	27			MyST_55h, PFS_10h	**7.94**	**2.91**	15.97	7.64
	28			LS_960h, MyST_55h, PFS_10h	8.13	3.12	**13.76**	**1.20**
	-	*Average (Group – E, BASE)*			*9.91*	*5.48*	*14.79*	*2.87*
	29	LARGE	Librilight	LS_960h, MyST_55h	8.06	9.31	13.20	1.34
	30			LS_960h, PFS_10h	13.18	3.17	13.19	**1.32**
	31			MyST_55h, PFS_10h	**7.42**	**2.99**	14.18	5.79
	32			LS_960h, MyST_55h, PFS_10h	8.17	3.33	**12.77**	1.40
	-	*Average (Group – E, LARGE)*			*9.2*	*4.7*	*13.33*	*2.4*

1) GROUP-A (TABLE-2)

In this group, adult datasets are used in both pretraining and finetuning. All models show a pattern of decreasing WER with an increase in the size of the finetuning dataset. It can also be observed that there is not a large difference in WER between BASE and LARGE models even though the LARGE model uses 60 times more training data.

2) GROUP-B (TABLE-3)

All the models in Group-B, finetuned with different amounts of MyST data, attained lower WERs on the child speech in comparison with Group-A experiments. A similar trend of decreasing WER can be observed with an increase in finetuning data.

3) GROUP-C (TABLE-3)

Group-C experiments were designed similar to Group-B (see Table-3). The objective was to investigate whether adding child speech dataset in the pretraining have any impact on the model performance. Comparing to the BASE models from Group-B, the WERs on all test sets increased in Group-C. Therefore, using child speech in pretraining was not considered for Group-D and Group-E experiments.

4) GROUP-D (TABLE-4)

In this group, the PFSTAR dataset was used for fine-tuning. The model's performance also improves as the size of the finetuning dataset increases. The best results, as might be expected, are on PFS_test while results on the other test datasets are less impressive.

5) GROUP-E (TABLE-5)

Group-E used LS_960h, PFS_10h and MyST_55h in various finetuning combinations as these datasets gave the best WER in previous finetuning experiments. Group-E models outperformed all the previous models and gave the best WER for all the inference datasets.

B. DISCUSSION OF RESULTS

Group-A (Table 2) results provide a baseline where only adult speech data is used for pretraining and finetuning. The relative improvements due to finetuning with adult speech are similar across all of the child test datasets, indicating that large adult speech datasets provide similar levels of improvement on different child speech validation. We can draw three additional conclusions. Firstly, there is less than a 3% variation in WER between BASE and LARGE wav2vec2 models across all the test datasets, so the LARGE model is only useful where optimal performance is needed, and BASE models are ideal for low resources scenario. Secondly, the improvement between finetuning with 10 minutes of adult speech data and 100 hours is much more significant than the improvement between 100 hours and 960 hours. There is only a 3% average WER difference between LS_100h and LS_960h finetuning, suggesting 100 hours of adult speech is ideal for finetuning.

Next, after introducing various amounts of child speech data for fine-tuning in Group-B (Table 3), it is noted that smaller amounts of child speech data result in better improvements in WER. It is clear that as little as 1 hour of child speech can have similar improvements to 100 hours of adult speech. Similarly, 10 hours of child speech shows similar improvements as 960 hours of adult speech. However, we also note a significant domain mismatch across the test datasets as the improvements on PFS_test and CMU_Kids are significantly weaker than for MyST_test. An overarching conclusion here might be that 1 hour of child speech is equivalent to 100 hours of adult speech where there is strong domain alignment between the finetuning and test speech. Lastly, using LARGE model for finetuning with only a small amount of child speech (e.g., 10 mins) may be detrimental due to domain mismatch between pretraining and finetuning datasets. Again, there is a relatively small performance improvement between BASE and LARGE models.

The Group-C (Table 3) experiments add the MyST_Complete dataset to the pretraining. Performance is poorer than with adult speech only, highlighting the limitations of pretraining data with the noisy and non-linguistic child speech in the MyST_Complete corpus. Further investigation is needed to understand this impact of child speech data on the pretraining; however, it will require a much cleaner and larger child speech dataset.

Group-D (Table 4) experiments are equivalent to Group-B (Table 3) but use the PFSTAR dataset for fine-tuning. As this dataset is smaller than MyST, only 10 minutes, 1 hour and 10 hours of speech can be used for fine-tuning. The key takeaway here is that PFS_test results improve even more significantly than MyST_test in Group-B, but the other child speech test datasets barely show any improvement. Clearly there is a significant domain mismatch between PFSTAR dataset with British English dialect and the two other child-speech datasets with American English dialect. PFSTAR was also recorded in a much cleaner environment. This shows that properties like dialect, accent and acoustic characteristics can impact the performance of the ASR model. Interestingly, MyST and PFSTAR finetuning (from Group B and D) shows similar WER on LibriTTS dev-clean implying that child speech datasets with distinct properties perform similarly when used for adult speech recognition.

Finally, for the Group-E experiments (Table 5), where a mix of adult and child datasets are used, we find that finetuning on the two child speech datasets, MyST_55h and PFS_10h gives the best results with WER rates of 7.91 and 2.94 on the respective tests datasets, MyST_test and PFS_test. Performance for CMU_Kids is significantly weaker at 15.97. Clearly, when the finetuning data has a good domain match with the tests data then SOTA WER rates can be achieved through finetuning with approximately 65 hours of child speech data. The BASE and LARGE configurations in Group-E show an absolute difference of 0.84 WER suggesting that performance is similar for both configurations when cross-domain datasets combinations are used in finetuning.

Interestingly, using smaller amounts of child speech can provide significant improvements in WER accuracy as compared with large amount of adult speech. This study provides a baseline for future studies. While the results of this study provide a comprehensive analysis of different

TABLE 6. Previous SOTA results on the MYST, PF-STAR, and CMU_KIDS datasets.

SOTA Papers	Method Type	Training Data (hrs)	Inference data (hrs)	WER MyST	WER PFSTAR	WER CMU_Kids
TDNN-F + Augmentation [30]	Supervised	6.34	2.76	-	-	16.01
Hybrid HMM-DNN Transfer Learning [28]	Supervised	6.26	2.45	-	-	19.33
DRAFT [37]: ○ WAV2VEC2 ○ HuBERT	Self-Supervised	197	13	 16.70 16.53	-	-
Transformer + CTC + Greedy [29]	Supervised	197	13	16.01	-	-
W2V2 + source-filter warping + LM [35]	Self-Supervised	11.2	2.5		4.86	

finetuning techniques for child ASR, additional conclusions can be drawn by comparing different experiments.

C. THIS WORK IN THE CONTEXT OF PREVIOUS CHILD SPEECH ASR APPROACHES

As commented in the Introduction, the publicly available child speech datasets are small in comparison to well-established adult speech datasets and audio quality is poor in comparison. Further, if the full datasets are used to build randomized test datasets, then many of the data samples will be of very variable quality. Thus, previous authors have adopted various approaches to clean and utilize the data but due to lack of standardized approach, it would not be fair to make any direct comparisons.

Our best results using the SSL approach show potential for significant improvement over the previously reported results on the same dataset as shown in Table 6. Our trained models achieved the best WER of **7.42** on the MyST_test dataset, **2.91** on the PFSTAR, PFS_test dataset (reaching human level performance) and **12.77** on the CMU_Kids dataset, as compared to the previously reported results from [28], [29], [30], [35], and [37]. Our detailed explanations of how the test datasets were 'cleaned' for this work should further provide researchers with a useful basis for future comparisons.

VI. CONCLUSION

In this work, the wav2vec2 self-supervised training approach is adapted with different mixes of pretraining and finetuning datasets to provide a methodology to improve the accuracy of child speech recognition. A combination of adult and child speech datasets is used to determine the data requirements for improving child speech recognition. Experiments were designed to evaluate the relative performance on the in-domain MyST and PFSTAR datasets, the out-of-domain CMUKIDS dataset while using the LibriTTS dev-clean dataset as a reference adult speech dataset. The best results were obtained where the model was pretrained on adult data and fine-tuned on a combinations of child speech datasets. The best WER rates (7.42 on MyST_test, 2.91 on PFS_test, 12.77 on CMU_Kids) are comparable with the best SOTA results available currently in the literature.

A model pretrained with adult speech data can best learn the speech features as compared to a model including both adult and child speech in pretraining. In particular, adding a low-quality dataset such as the MyST child speech dataset in pretraining reduced the performance of the ASR model across all test datasets. Significant domain variations were also evident between the MyST, CMU_Kids and PFSTAR datasets with the latter being of notably better quality. Qualitatively we can say that MyST and CMU_Kids are more closely aligned than the PF-Star dataset. When a cross-domain mix of child speech is used for fine-tuning (e.g., model 27 or model 31) then the optimal results are achieved. For a model finetuned with single or multiple child/adult speech data, WER increases over the dataset with similar distribution as finetuning dataset.

The BASE configuration of wav2vec2, which is pretrained with 60 times less data than the LARGE configuration is effective for a low-data scenario. In fact, the improvements achieved through using the LARGE configuration were typically only a few percent and hardly seem to justify the large increase in computational resources needed to train. As for finetuning, we can say that 100 hours of adult speech finetuning data offer a practical trade-off between computational effort and ASR accuracy. Finetuning with as little as 10 hours of child speech data provided better improvement over models finetuned with 960 hours of adult speech. Optimal results are achieved using in the order of 65 hours of cross-domain child speech (a mix of MyST and PFSTAR).

For future work, these models can be used to transcribe additional child speech data from the unlabeled MyST dataset and a range of additional unlabeled datasets. It would also be interesting to investigate the potential of generative data augmentation models [47] to provide additional synthetic child speech samples and a wider variety of child speech for pretraining and finetuning experiments.

ACKNOWLEDGMENT

The authors would like to thank the experts from Xperi Ireland: Gabriel Costache, Zoran Fejzo, George Sterpu and the rest of the team members for providing their expertise and feedback throughout.

REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, and J. Chen, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[3] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*.

[4] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6124–6128.

[5] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[6] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "ContextNet: Improving convolutional neural networks for automatic speech recognition with global context," 2020, *arXiv:2005.03191*.

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[8] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999.

[9] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 1997, pp. 1–4.

[10] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about databases of children's speech," in *Proc. Interspeech*, Aug. 2013, pp. 2410–2414.

[11] V. Bhardwaj, M. T. B. Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (ASR) systems for children: A systematic literature review," *Appl. Sci.*, vol. 12, no. 9, p. 4419, Apr. 2022, doi: 10.3390/APP12094419.

[12] R. Peinl and J. Wirth, "Quality assurance for speech synthesis with ASR," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2022, pp. 739–751.

[13] A. Baby, S. Vinnaitherthan, N. Adiga, P. Jawale, S. Badam, S. Adavanne, and S. Konjeti, "An ASR guided speech intelligibility measure for TTS model selection," Jun. 2020, *arXiv:2006.01463*.

[14] V. Kadyan, H. Kathania, P. Govil, and M. Kurimo, "Synthesis speech based data augmentation for low resource children ASR," in *Speech and Computer* (Lecture Notes in Computer Science), vol. 12997. Cham, Switzerland: Springer, 2021, pp. 317–326, doi: 10.1007/978-3-030-87802-3_29.

[15] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 976–989, 2020, doi: 10.1109/TASLP.2020.2977776.

[16] S. Shahnawazuddin, N. Adiga, H. K. Kathania, and B. T. Sai, "Creating speaker independent ASR system through prosody modification based data augmentation," *Pattern Recognit. Lett.*, vol. 131, pp. 213–218, Mar. 2020, doi: 10.1016/j.patrec.2019.12.019.

[17] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian, "Towards data selection on TTS data for children's speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6888–6892, doi: 10.1109/ICASSP39728.2021.9413930.

[18] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario," in *Proc. Interspeech*, 2020, pp. 4382–4386, doi: 10.21437/Interspeech.2020-1112.

[19] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, "Data augmentation using CycleGAN for end-to-end children ASR," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 511–515, doi: 10.23919/EUSIPCO54536.2021.9616228.

[20] N. Jia, C. Zheng, and W. Sun, "Speech synthesis of children's reading based on CycleGAN model," *J. Phys., Conf.*, vol. 1607, no. 1, Aug. 2020, Art. no. 012046, doi: 10.1088/1742-6596/1607/1/012046.

[21] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2014, pp. 135–140.

[22] G. Yeung, R. Fan, and A. Alwan, "Fundamental frequency feature normalization and data augmentation for child speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6993–6997, doi: 10.1109/ICASSP39728.2021.9413801.

[23] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *Proc. Interspeech*, Sep. 2016, pp. 1598–1602, doi: 10.21437/INTERSPEECH.2016-1348.

[24] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition—The 'ethiopian' system for the SLT 2021 children speech recognition challenge," Nov. 2020, *arXiv:2011.04547*.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.

[26] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral modification based data augmentation for improving end-to-end ASR for children's speech," 2022, *arXiv:2203.06600*.

[27] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Comput. Speech Lang.*, vol. 63, Sep. 2020, Art. no. 101077, doi: 10.1016/J.CSL.2020.101077.

[28] T. Rolland, A. Abad, C. Cucchiarini, and H. Strik, "Multilingual transfer learning for children automatic speech recognition," in *Proc. 13th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 7314–7320.

[29] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101289.

[30] F. F. Wu, L. P. Garcia, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network," in *Proc. Interspeech*, 2019, pp. 1–5, doi: 10.21437/Interspeech.2019-2980.

[31] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 27826–27839.

[32] G. Xu, S. Yang, L. Ma, C. Li, and Z. Wu, "The TAL system for the INTERSPEECH2021 shared task on automatic speech recognition for non-native children's speech," in *Proc. Interspeech*, 2021, pp. 1294–1298, doi: 10.21437/Interspeech.2021-1104.

[33] F. Yu, Z. Yao, X. Wang, K. An, L. Xie, Z. Ou, B. Liu, X. Li, and G. Miao, "The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 1117–1123.

[34] R. Fan, A. Afshan, and A. Alwan, "Bi-APC: Bidirectional autoregressive predictive coding for unsupervised pre-training and its application to children's ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7023–7027.

[35] J. Thienpondt and K. Demuynck, "Transfer learning for robust low-resource children's speech ASR with transformers and source-filter warping," 2022, *arXiv:2206.09396*.

[36] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards better domain adaptation for self-supervised models: A case study of child ASR," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1242–1252, Oct. 2022, doi: 10.1109/JSTSP.2022.3200910.

[37] R. Fan and A. Alwan, "DRAFT: A novel framework to reduce domain shifting in self-supervised learning and its application to children's ASR," in *Proc. Interspeech*, 2022, pp. 1–5, doi: 10.21437/Interspeech.2022-11128.

[38] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[39] E. Jang, G. Brain, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. ICLR*, 2017, pp. 1–13.

[40] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

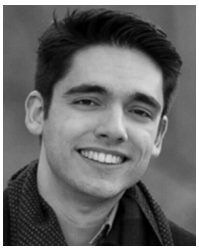
[41] W. Ward, R. Cole, and S. Pradhan, "My science tutor and the myst corpus," Boulder Learn. Inc., 2019.

[42] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids corpus LDC97S63," Tech. Rep. LDC97S63, 1997. [Online]. Available: https://catalog.ldc.upenn.edu
[43] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," 2005.
[44] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for ASR with limited or no supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7669–7673.
[45] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech*, 2019, pp. 1–7, doi: 10.21437/interspeech.2019-2441.
[46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.
[47] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis," *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.

RISHABH JAIN (Graduate Student Member, IEEE) received the B.Tech. degree in computer science and engineering from the Vellore Institute of Technology (VIT), in 2019, and the M.S. degree in data analytics from the University of Galway, Ireland, in 2020, where he is currently pursuing the Ph.D. degree. He is a Research Assistant with the University of Galway under the Data-center Audio/Visual Intelligence on-Device (DAVID)Project. His research interests include machine learning and artificial intelligence specifically in the domain of speech understanding, text-to-speech, speaker recognition, and automatic speech recognition.

ANDREI BARCOVSCHI received the B.Eng. degree in electronic and computer engineering from the University of Galway (prior to 2023: National University of Ireland Galway (NUIG)), in 2020 and the M.Sc. degree in artificial intelligence from NUIG, in 2021. He is currently pursuing the Ph.D. degree in artificial intelligence with the University of Galway. His research interests include speech synthesis and conversion technologies, text-to-speech, and speech-to-text. He is interested in a broad range of machine learning and artificial intelligence topics.

MARIAM YAHAYAH YIWERE received the B.Sc. degree from the Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, in 2012, and the M.Eng. and Ph.D. degrees from the Department of Computer Engineering, Hanbat National University, South Korea, in August 2015 and February 2020, respectively. Since October 2020, she has been working on the DTIF/DAVID Project, as a Postdoctoral Researcher with the College of Science and Engineering, University of Galway, Ireland. Her research interests include text-to-speech synthesis, speaker recognition and verification, sound source localization, deep learning, and computer vision.



DAN BIGIOI (Graduate Student Member, IEEE) received the bachelor's degree in electronic and computer engineering from the National University of Ireland Galway, in 2020. He is currently pursuing the Ph.D. degree with the University of Galway, sponsored by D-REAL, the SFI Centre for Research Training in Digitally Enhanced Reality. Upon graduating, he worked as a Research Assistant with the University of Galway, Ireland, studying the text-to-speech and speaker recognition methods under the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project. His research interests include novel deep learning-based techniques for automatic speech dubbing and discovering new ways to process multi-modal audio/visual data.

PETER CORCORAN (Fellow, IEEE) is currently the Personal Chair of electronic engineering with the College of Science and Engineering, University of Galway, Ireland. He was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He has more than 600 cited technical publications and patents, more than 120 peer-reviewed journal articles, and 160 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is a member of the IEEE Consumer Technology Society for more than 25 years. He is the founding Editor of *IEEE Consumer Electronics Magazine*.

HORIA CUCU (Member, IEEE) received the B.S. and M.S. degrees in applied electronics and the Ph.D. degree in electronics and telecom from the University Politehnica of Bucharest (UPB), Romania, in 2008 and 2011, respectively.

From 2010 to 2017, he was a Teaching Assistant and a Lecturer with UPB, where he is currently an Associate Professor. In this position, he authored more than 75 scientific papers in international conferences and journals, served as the Project Director for seven research projects, and contributed as a Researcher to ten other research grants. He holds two patents. In addition, he founded and leads Zevo Technology, a speech start-up dedicated to integrating state-of-the-art speech technologies in various commercial applications. His research interests include machine/ deep learning and artificial intelligence, with a special focus on automatic speech and speaker recognition, text-to-speech synthesis, and speech emotion recognition.

Dr. Cucu was awarded the Romanian Academy prize "Mihail Drăgănescu", in 2016, for outstanding research contributions in Spoken Language Technology, after developing the first large-vocabulary automatic speech recognition system for the Romanian language.

Appendix C

Adaptation of Whisper models to child speech recognition.

Authors: Rishabh Jain (RJ), Andrei Barcovschi (AB), Mariam Yiwere (MY), Peter Corcoran (PC) and Horia Cucu (HC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	RJ: 80%, AB: 5%, MY: 5%, PC:10%
Experiments and Implementation	RJ: 90%, AB: 10%
Background	RJ: 80%, AB: 10%, MY: 10%
Manuscript Preparation	RJ: 70%, AB: 10%, MY: 10%, PC: 5%, HC: 5%

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Adaptation of Whisper models to child speech recognition

Rishabh Jain[1], *Andrei Barcovschi*[1], *Mariam Yiwere*[1], *Peter Corcoran*[1], *Horia Cucu*[2]

[1]School of Electrical and Electronics Engineering, University of Galway, Galway, Ireland
[2] Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Romania
rishabh.jain@universityofgalway.ie, a.barcovschi1@universityofgalway.ie,
mariam.yiwere@universityofgalway.ie, peter.corcoran@universityofgalway.ie,
horia.cucu@upb.ro

Abstract

Automatic Speech Recognition (ASR) systems often struggle with transcribing child speech due to the lack of large child speech datasets required to accurately train child-friendly ASR models. However, there are huge amounts of annotated adult speech datasets which were used to create multilingual ASR models, such as Whisper. Our work aims to explore whether such models can be adapted to child speech to improve ASR for children. In addition, we compare Whisper child-adaptations with finetuned self-supervised models, such as wav2vec2. We demonstrate that finetuning Whisper on child speech yields significant improvements in ASR performance on child speech, compared to non-finetuned Whisper models. Additionally, utilizing self-supervised Wav2vec2 models that have been finetuned on child speech outperforms Whisper finetuning.

Index Terms: Child Speech Recognition, Automatic Speech Recognition, Whisper model, MyST, PF-STAR, CMU Kids

1. Introduction

Automatic Speech Recognition (ASR) faces several challenges, including limited training data, untranscribed training data and performance degradation on non-native speech and children's speech. Recent research in ASR tackles some of these problems, especially for adult speech, and therefore ASR on adult speech has reached human-level performance [1]–[4]. However, for child speech, progress has been slow and ASR models still perform poorly. Unlike adult speech data, high quality child speech datasets required for training are limited and challenging to collect and annotate (see the survey in [5]). Additionally, there are inherent differences between adult and child voices in terms of pitch, linguistic and acoustic features, and pronunciation ability [6], [7]. The shorter vocal tract length and higher fundamental frequency [8] of children's voices also add to the complexity of recognizing child speech.

Recent development in self-supervised learning has delivered improvements for child speech. The development of unsupervised pretraining techniques, such as Wav2vec2 [3], has greatly contributed to the progress of child ASR [9]–[11]. However, a finetuning stage on a labeled dataset is required for ASR, which limits their usefulness since finetuning can find patterns within a training dataset and boost performance on the similar datasets but may not generalize to other dataset distributions. The aim of speech recognition systems is to operate with high reliability in diverse environments, without the need for finetuning for the data/deployment distribution of each specific usecase. We reviewed various supervised learning approaches [12]–[14] in child ASR. It was observed that most of these studies included transfer learning approaches from adult to child speech [9], [12], [15], data augmentation methods [16]–[20], or weakly supervised training [14], [15], [21]. Recent findings in supervised learning approaches [22], [23] has demonstrated that pretraining speech recognition models on multiple datasets/domains using supervised methods can enhance the models' robustness and generalization performance on unseen datasets.

In this work, we use a recent State-of-the-Art (SOTA) supervised ASR model, called Whisper. The authors of Whisper [4] have successfully bridged the gap in weakly supervised speech recognition by using large amounts of labeled audio data. They have also broadened the scope of weakly supervised pre-training beyond English-only speech recognition to be multilingual and multitask, showing great performance on different multilingual adult speech datasets [4]. These findings suggest that the scaling of weakly supervised pretraining has been undervalued for speech recognition. We use these Whisper models to provide an analysis of supervised training paradigms on different child speech datasets. We also finetune these models using different combinations of child speech datasets to see the subsequent speech recognition performance on different seen and unseen distributions of child speech datasets [24]–[26]. Lastly, we provide a comparative analysis of Whisper results with previously benchmarked results that used wav2vec2 self-supervised learning approach trained on the same distribution of datasets [27]. We use a similar approach as used by the authors of [28] for providing a comparison between Whisper and wav2vec2 results.

Since Whisper is trained with an order of magnitude more data than wav2vec2 (680k vs 60k) and contains a lot of multilingual and low resource languages during training, we believe that this multilingual data can be utilized to provide child speech recognition tasks via finetuning. Our goal is to evaluate the efficacy of these two methodologies in child speech analysis and determine their potential for enhancing child ASR technology and developing educational tools for children.

2. Model Description

2.1. **Whisper** [4]

The Whisper approach focuses on broadening the scope of weakly supervised pre-training beyond English-only speech recognition to be both multilingual and multitask. Of the 680,000 hours of labelled audio used by Whisper, 117,000 hours cover 96 other languages. The dataset also includes 125,000 hours of X→en translation data. The model processes audio through a system of transformer blocks with residual

10.21437/Interspeech.2023-935

connections and final layer normalization. The model uses a multitask format to perform the entire speech processing pipeline, including transcription, translation, voice activity detection, alignment, and language identification. The model is based on an encoder-decoder Transformer, which is fed 80-channel log-Mel spectrograms. The encoder is formed by two convolutional layers with a kernel size of 3, followed by a sinusoidal positional encoding, and a stacked set of Transformer blocks. The decoder uses the learned positional embeddings and the same number of Transformer blocks as the encoder. The Whisper architecture is explained in detail in [4].

2.2. **Wav2vec2** [3]

Wav2vec 2.0 is a speech recognition model and training approach that is based on a self-supervised learning of speech representations using a two-stage architecture for pretraining and finetuning. The architecture of wav2vec 2.0 can be divided into three main parts: a CNN feature extractor, a transformer-based encoder, and a quantization module (see [3] for more details) . In the pretraining phase, the model is trained on a large dataset of unlabelled speech data. The model learns meaningful representations by capturing the temporal and spectral characteristics of speech using a masked contrastive loss function. In the finetuning phase, the pretrained model is finetuned on a smaller labeled dataset for a specific downstream task. The last layer of the pretrained model is replaced with a task-specific feed-forward layer and the entire model is optimized by minimizing the CTC loss [29] for ASR.

2.3. Training details

All models were trained using A6000 GPUs with 48GB of available memory. We provide the architectural parameters details in Table 1 for both Whisper and wav2vec2 models used in this work. Whisper models are trained with a large number of parameters and therefore should provide better generalization towards unseen datasets compared to wav2vec2.

Table 1: *Architecture parameters for Whisper* [4] and wav2vec2 [3] models.

Models	Layers	Width	Heads	Learning Rate	Para-meters
Whisper Models:					
Tiny	4	384	6	1.5×10^{-3}	39M
Base	6	512	8	1×10^{-3}	72M
Small	12	768	12	5×10^{-4}	244M
Medium	24	1024	16	2.5×10^{-4}	769M
Large	32	1280	20	1.75×10^{-4}	1550M
Wav2vec2 Models:					
Base	12	768	8	5×10^{-4}	95M
Large	24	1024	16	3×10^{-4}	317M

For finetuning, we use a learning rate of 1×10^{-5} for all Whisper finetuning experiments. Wav2vec2-base was finetuned with a learning rate of 1×10^{-4} , while wav2vec2-large was finetuned with a learning rate of 2.5×10^{-5}, consistent with [3]. Finetuning both approaches involve training the final layer of the models and freezing all others, as described by the respective authors. Finetuning parameters were kept the same as provided in Whisper [4] and wav2vec2 [3]. The Whisper model undergoes finetuning by minimizing the cross-entropy objective function, whereas wav2vec2 is finetuned by minimizing the CTC loss.

3. Corpus Description

The authors of Whisper [4] do not mention the datasets used. However, these trained models achieved SOTA results on many different adult speech ASR datasets [4]. For our work, we use three different child speech datasets and one adult speech dataset: MyST Corpus [24], PFSTAR dataset [25], CMU Kids dataset [26] and LibriTTS dev-clean dataset [30]. The datasets are kept consistent with previous research [27] on wav2vec2 to provide objective comparison with the Whisper models.

3.1. Dataset Cleanup

All the labeled data was cleaned as per the guidelines mentioned by the authors of Whisper [4]. The abbreviations, punctuations, white spaces, and other non-alphanumeric characters were removed, and all the characters were changed to lowercase. Audio data was modified to have a 16Khz sampling rate and be 16-bit mono channel. The 'dev-clean' subset of LibriTTS [30], containing 9 hours of audio is used to provide an evaluation of our experiments on adult speech. My Science Tutor (MyST) Corpus [24] is an American English child speech dataset containing over 393 hours of child speech, of which 197 hours are fully transcribed. The dataset was cleaned and prepared as mentioned in [27], with 65 hours of clean child speech divided into two subsets: 55 hours for training and 10 hours of testing. PFSTAR [25] includes a collection of words spoken by British English children and contains a total of 12 hours of audio. 10 hours of this data was used for training and 2 hours was held out for inference. CMU Kids [26] corpus was used for validation-only, which contains 9 hours of read-aloud sentences by children recorded at Carnegie Mellon University. While these may not be very big speech datasets, they currently represent the best publicly available child speech datasets.

3.2. Dataset Usage

The datasets were divided according to their usage for 'training' and 'inference'. This information is summarized in Table 2.

Table 2: *Dataset usage*

Usage	Dataset	Duration
Finetuning (Training)	MyST_55h	55 hours
	PFS_10h	10 hours
Inference (Testing)	dev-clean	9 hours
	MyST_test	10 hours
	PFS_test	2 hours
	CMU_test	9 hours

4. Experiments and Results

4.1. Codebase

The Whisper implementation used is provided here [1] . The fairseq [2] implementation of wav2vec2 is used for finetuning experiments. Our trained Whisper models are available to use on the HF platform [3]. The relevant information regarding model training, hyperparameters, graphs/metrics, checkpoints, and dataset availability are made available on our GitHub [4].

[1] **Whisper Implementation:** https://github.com/huggingface/community-events/tree/main/whisper-fine-tuning-event
[2] **Wav2vec2 Fairseq:** https://github.com/facebookresearch/fairseq/
[3] **Finetuned Whisper models:** https://huggingface.co/rishabhjain16
[4] **GitHub:** https://github.com/C3Imaging/whisper_child_speech

4.2. Experiments

In our first set of experiments (see Section 4.3.1), the original Whisper models were evaluated on different child speech datasets mentioned in Table 2. The models are categorized based on their size: Tiny, Base, Small, Medium, Large, and Large V2 (see Table 1). 'Large-V2' was trained for 2.5X more epochs as compared to 'Large', while also adding extra parameters for regularization [4]. There are two versions of each model: one trained with multilingual data and one specifically for the English language only (indicated by '.en' in the name). 'Large' and 'Large-V2' models don't have English-only models. Figure 1 shows a plot comparing Word Error Rate (WER) on 12 English adult speech datasets against model parameters (as provided by Whisper[4]). As expected, lower WER values are obtained using models with more parameters. We also perform a similar comparison using our child speech datasets (more in section 4.3).

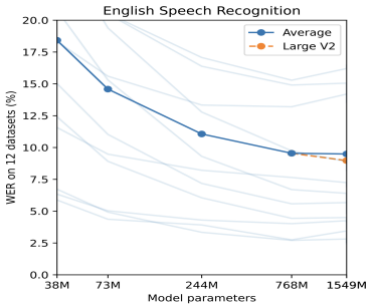


Figure 1: *Whisper Parameters vs. WER on adult speech datasets (from* [4]).

The second set of experiments (see Section 4.3) involved finetuning these Whisper models with child speech. Three models with the best performance from the first set of experiments are selected for further finetuning. We finetuned each of the selected models up to 4000 epochs. We select the best performing checkpoints from among the trained models, which shows the lowest WER while training. Finetuning included three experimental configurations of training data: MyST_55h, PFSTAR_10h, and MyST_55h+PFSTAR_10h combined. These finetuning experiments were kept consistent with previously reported wav2vec2 finetuning experiments [27] in order to compare both models trained with a similar distribution of finetuning data. The wav2vec2 'base' and 'large' models are used for finetuning, which are pretrained with 960 hours of Librispeech data [31], and 60,000 hours of Librilight data [32], respectively. The difference in their parameters sizes can be seen in Table 1. This comparison is provided to see how supervised and self-supervised approaches behave with child speech.

4.3. Results and Discussion

4.3.1. Whisper Original (No-Finetuning):

Table 3 provides the WER results on the inference datasets using different original Whisper models from the first set of experiments. These models are provided by the authors [4] and no initial finetuning was performed over these models. It can be observed that the models with larger numbers of parameters generally perform better. Among the models with the same number of parameters, the English models perform better than the multilingual models, suggesting that training on language-specific data can improve performance for that language. The lowest WER achieved are highlighted in Table 3.

Table 3: *WER for different Whisper and Wav2vec2 models (without finetuning) on child speech (MyST, PFSTAR and CMU Kids) and adult speech (dev-clean) datasets.*

Models	MyST_ test	PFS_ test	CMU_ test	dev-clean
Tiny	40.09	159.57	30.63	10.85
Tiny.en	33.02	47.11	27.32	8.62
Base	32.14	100.07	25.03	8.14
Base.en	29.15	45.70	20.75	7.18
Small	26.22	111.75	18.52	6.43
Small.en	26.72	39.00	16.82	6.06
Medium	25.11	80.97	**12.67**	5.58
Medium.en	28.06	**35.25**	14.00	6.20
Large	25.24	84.52	13.70	5.53
Large-V2	**25.00**	73.68	12.69	**5.40**
w2v2-base (LS_960)	15.41	11.20	16.33	3.40
w2v2-large (LL_60k)	**12.50**	**8.56**	**14.85**	**3.28**

Note: '.en' respresents the English-only trained models, while all others represent the multilingual models. For example, 'Tiny' contains both English and other multilingual training data while 'Tiny.en' contains only English speech. Wav2vec2 results presented for comparison are taken from previously presented work on wav2vec2 for child ASR [27]. The 'w2v2-base' is pretrained with 960 hours of Librispeech data (LS_960) and 'w2v2-large' is pretrained with 60k hours of Librilight data (LL_60k). Both models were finetuned using Librispeech for providing a comparison with non-finetuned Whisper models. The WER reported in Table 3 uses zero-shot setting.

These models achieved positive results on multilingual adult speech without the need to perform data-specific finetuning (see Figure 1), however, the performance seems poor for child speech, despite Whisper stating that their models generalize well to standard benchmarks in a zero-shot transfer setting without the need for any finetuning. We use these experiments as a baseline for further finetuning. The models with lowest WER were chosen ('Medium', 'Medium.en' and 'Large-V2') for providing further finetuning with child speech.

4.3.2. Whisper Finetuning with Child Speech

The Whisper finetuning experiments include three subsets of experiments: finetuning with MyST_55h, PFSTAR_10h and a combination of both datasets. Table 4 shows the WER of the selected finetuned models using these subsets. During finetuning, cross entropy loss is minimized by training only on the last layer and freezing all other layers, allowing the model to classify target tokens from a predefined vocabulary.

Table 4: *WER on inference (test) datasets for different Whisper and wav2vec2 models finetuned on MyST, PFSTAR and MyST+PFSTAR-combined datasets.*

ID	Models	MyST_ test	PFS_ test	CMU_ test	dev-clean
MyST (55 Hours) Finetuning:					
1	Medium	**11.66**	19.76	16.84	5.62
2	Medium.en	11.81	17.83	**15.07**	6.48
3	Large-V2	12.28	**10.88**	15.67	**4.82**
4	w2v2-base	8.13	14.77	16.47	7.72
5	w2v2-large	**7.51**	**12.46**	**15.25**	**6.43**
PFSTAR (10 Hours) Finetuning:					
6	Medium	16.18	3.15	16.57	5.33
7	Medium.en	15.84	3.14	15.53	5.28
8	Large-V2	**15.79**	**2.88**	**15.22**	**5.10**
9	w2v2-base	31.86	**3.48**	27.49	13.95
10	w2v2-large	**27.17**	3.50	**21.35**	**11.60**
MyST (55 Hours) + PFSTAR (10 Hours) Finetuning:					
11	Medium	**12.22**	**2.98**	16.05	5.40
12	Medium.en	12.33	3.32	**15.08**	**4.88**
13	Large-V2	13.34	4.17	17.11	4.97
14	w2v2-base	7.94	**2.91**	15.97	7.64
15	w2v2-large	**7.42**	2.99	**14.18**	**5.79**

Note: Wav2vec2 results are taken from [27]. The 'w2v2-base' represents wav2vec2 base model while 'w2v2-large' represents wav2vec2 large models.

Finetuning with MyST_55h showed a significant improvement in the WER of MyST_test and PFS_test. However, CMU_test dataset had a 2% increase in WER, as shown in Table 4. WER on dev-clean adult speech dataset also decreased by 1%. Finetuning with PFS_10h also had a significant improvement on MyST_test and PFS_test. The WER on both test sets decreased; however, the improvement in WER on the MyST_test is not as good as when the models are finetuned with MyST_55h. CMU_test had a 2% increase in WER, similar to MyST_finetuning. Large-V2 Whisper model gave the lowest WER on all four inference data setups, with WER on PFS_test dropping to 2.88. When both MyST_55h and PFS_10h were used for finetuning, the WER on both MyST_test and PFS_test dropped significantly. It can be observed that for a dataset used in finetuning, the model shows an improvement in performance on datasets with similar distribution at inference time.

The following observations were seen in all finetuning experiments: Whisper finetuned models yield better results than Whisper original models, regardless of dataset distribution, but a finetuning dataset that matches the distribution of the test dataset can improve performance. CMU_test showed an increase in WER regardless of the finetuning setup and remained in the range of 15-17%. This could imply that CMU Kids might be a noisy dataset which doesn't work well for ASR. The WER of dev-clean adult speech further decreased after child speech finetuning and stayed in the range of 4-5% for all experiments.

4.3.3. Whisper vs Wav2vec2:

We compare Whisper models with wav2vec2 finetuned models on the same datasets. Table 3 and Table 4 cover the various wav2vec2 finetuning results on different child speech datasets. We first compare Librispeech-finetuned 'base' and 'large' wav2vec2 models with the original Whisper 'Medium' and 'Large' models (See Table 3). This was done to maintain consistency with the comparison mechanism as provided by authors of Whisper [4]. The wav2vec2 models finetuned with Librispeech generally performed better on child speech compared to any of the Whisper models without finetuning. Both these models were used to provide a usecase of ASR over unseen child speech in low resource data scenario. Wav2vec2 results show the lowest WER on all inference datasets except CMU_test. However, Whisper models gave lower WER on CMU_test as compared to wav2vec2 models. This implies that CMU kids dataset could have acoustic properties similar to adult speech since supervised finetuning using Whisper decreases the WER on CMU_test.

The results of the experiments with child speech finetunings show that wav2vec2 finetuning using MyST_55h resulted in lower WER compared to Whisper finetuning on MyST_test. However, an increase in WER was observed on PFS_test and dev-clean for wav2vec2 finetuning. Both Whisper and wav2vec2 finetuned models had a WER range of 14-16% on CMU_test. For PFS_10h finetuning, similar results were obtained for both wav2vec2 and Whisper models on PFS_test, with WER of 3.48 and 2.88, respectively. However, high WERs were observed on all other inference datasets. These results suggest that wav2vec2 finetuning generalizes well for datasets with a similar distribution, while Whisper finetuning works best for unseen datasets at inference time. When both MyST_55h and PFS_10h were used for finetuning, the lowest WER was observed with wav2vec2 finetuning across all child speech datasets as compared to Whisper finetuning. Both Whisper and wav2vec2 models behaved similarly when finetuned with a combination of child speech datasets, but wav2vec2 performed better on datasets with similar distributions as the seen datasets. Moreover, when considering the amount of training data and model size (model 13 vs model 14), it was observed that the wav2vec2 model 15 (60k hours, 317M parameters) performed better than Whisper model 13 (680k hours, 1550M parameters), which were finetuned with the same amount of child speech data. These findings demonstrate that wav2vec2 performs well with child speech and slightly outperforms Whisper.

5. Conclusions

In this paper, we use the recent SOTA large-scale supervised Whisper models for experimental analysis over different child speech datasets. The study of different combinations of finetuning over child-specific datasets is also presented in this paper. Finetuning Whisper models achieved significant improvements in accuracy of child speech recognition. We also present comparisons with the SOTA self-supervised, wav2vec2 model. Finetuning both Whisper and wav2vec2 improves performance of child ASR. While Whisper improves ASR performance for both adult and child speech, regardless of the finetuning dataset, wav2vec2 model performs better with finetune-specific datasets. Although Whisper may be more appropriate for unseen datasets, wav2vec2 is a better choice for real-time, task-specific applications. In addition, the use of smaller-sized models, such as wav2vec2, would be more feasible for deployment on edge devices, which is also using 10 times less training data than Whisper. For future work, we aim to further study this methodology by including more low resource datasets (both adult and child), different ASR decoding strategies and deploying these models on edge devices.

6. Acknowledgements

The authors would like to acknowledge experts from Xperi Ireland: Gabriel Costache, Zoran Fejzo, and George Sterpu for providing their expertise and feedback while working on this research.

7. References

[1] Kriman, Samuel, et al. "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[2] Gulati, Anmol, et al. "Conformer: Convolution-augmented transformer for speech recognition." *arXiv preprint arXiv:2005.08100* (2020).

[3] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." *arXiv preprint arXiv:2212.04356* (2022).

[5] Claus, Felix, Hamurabi Gamboa Rosales, Rico Petrick, Horst-Udo Hain, and Rüdiger Hoffmann. "A survey about databases of children's speech." In *INTERSPEECH*, pp. 2410-2414. 2013.

[6] S. Lee, A. Potamianos, and S. Narayanan, “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters,” *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999, doi: 10.1121/1.426686.

[7] S. Lee, A. Potamianos, and S. S. Narayanan, “Analysis of children’s speech: duration, pitch and formants,” in *EUROSPEECH*, 1997.

[8] R. Serizel and D. Giuliani, “Vocal tract length normalisation approaches to DNN-based children’s and adults’ speech recognition,” pp. 135–140, 2014, doi: 10.1109/SLT.2014.7078563ï.

[9] Thienpondt, Jenthe, and Kris Demuynck. "Transfer Learning for Robust Low-Resource Children’s Speech ASR with Transformers and Source-Filter Warping." *Proc. Interspeech 2022*.

[10] Fan, Ruchao, Yunzheng Zhu, Jinhan Wang, and Abeer Alwan. "Towards better domain adaptation for self-supervised models: A case study of child asr." *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, 2022, doi: 10.1109/JSTSP.2022.3200910.

[11] R. Fan and A. Alwan, “DRAFT: A Novel Framework to Reduce Domain Shifting in Self-supervised Learning and Its Application to Children’s ASR,” 2022, doi: 10.21437/Interspeech.2022-11128.

[12] T. Rolland, A. Abad, C. Cucchiarini, and H. Strik, “Multilingual Transfer Learning for Children Automatic Speech Recognition,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Jun. 2022, pp. 7314–7320.

[13] F. Wu, L. Paola Garcia, D. Povey, and S. Khudanpur, “Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network,” 2019, doi: 10.21437/Interspeech.2019-2980.

[14] P. Gurunath Shivakumar and S. Narayanan, “End-to-end neural systems for automatic children speech recognition: An empirical study,” *Computer Speech & Language*, vol. 72, Mar. 2022, doi: 10.1016/j.csl.2021.101289.

[15] P. Gurunath Shivakumar and P. Georgiou, “Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations,” *Computer Speech & Language*, vol. 63, p. 101077, Sep. 2020, doi: 10.1016/J.CSL.2020.101077.

[16] K. Y. Chenpeng Du, “Speaker Augmentation for Low Resource Speech Recognition,” *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7719–7723, 2020.

[17] H. K. Kathania, V. Kadyan, S. R. Kadiri, and M. Kurimo, “Data Augmentation Using Spectral Warping for Low Resource Children ASR,” *Journal of Signal Processing Systems*, vol. 94, no. 12, pp. 1507–1513, Dec. 2022, doi: 10.1007/S11265-022-01820-0/TABLES/6.

[18] V. Kadyan, H. Kathania, P. Govil, and M. Kurimo, “Synthesis Speech Based Data Augmentation for Low Resource Children ASR,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12997 LNAI, pp. 317–326, 2021, doi: 10.1007/978-3-030-87802-3_29.

[19] G. Yeung, R. Fan, and A. Alwan, “Fundamental Frequency Feature Normalization and Data Augmentation for Child Speech Recognition,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6993–6997. doi: 10.1109/ICASSP39728.2021.9413801.

[20] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, “Data Augmentation Using CycleGAN for End-to-End Children ASR,” *European Signal Processing Conference*, vol. 2021-August, pp. 511–515, 2021, doi: 10.23919/EUSIPCO54536.2021.9616228.

[21] Gerosa, Matteo, et al. "A review of ASR technologies for children's speech." *Proceedings of the 2nd Workshop on Child, Computer and Interaction*. 2009.

[22] A. Narayanan *et al.*, “Toward Domain-Invariant Speech Recognition via Large Scale Training,” *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*, pp. 441–447, Feb. 2019, doi: 10.1109/SLT.2018.8639610.

[23] W. Chan, D. S. Park, C. A. Lee, Y. Zhang, Q. v Le, and M. Norouzi, “SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network.”

[24] Ward, Wayne, Ron Cole, and Sameer Pradhan. "My science tutor and the myst corpus." *Boulder Learning Inc* (2019).

[25] Russell, Martin. "The pf-star british english childrens speech corpus." *The Speech Ark Limited* (2006).

[26] M. Eskenazi, J. Mostow, and D. Graff, “The CMU kids speech corpus,” *Corpus of children's read speech digitized and transcribed on two CD-ROMs, with assistance from Multicom Research and David Graff. Published by the Linguistic Data Consortium, University of Pennsylvania*, 1997.

[27] R. Jain, A. Barcovschi, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, “A Wav2vec2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition,” Apr. 2022, doi: 10.48550/arxiv.2204.05419.

[28] S. Squartini, M. Scarpiniti, J.-T. Chien, J. Camilo Vásquez-Correa, and A. Á. Muniain, “Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper,” *Sensors 2023, Vol. 23, Page 1843*, vol. 23, no. 4, p. 1843, Feb. 2023, doi: 10.3390/S23041843.

[29] Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." *Proceedings of the 23rd international conference on Machine learning*. 2006.

[30] H. Zen *et al.*, “LibriTTS: A corpus derived from libri speech for text-to-speech,” *arXiv*. 2019. doi: 10.21437/interspeech.2019-2441.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.

[32] Kahn, Jacob, et al. "Libri-light: A benchmark for asr with limited or no supervision." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

Appendix D

Improved Child Text-to-Speech Synthesis through Fastpitch-based Transfer Learning.

Authors: Rishabh Jain (RJ) and Peter Corcoran (PC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	RJ: 90%, PC:10%
Experiments and Implementation	RJ: 100%
Background	RJ: 100%
Manuscript Preparation	RJ: 90%, PC:10%

Improved Child Text-to-Speech Synthesis through Fastpitch-based Transfer Learning

Rishabh Jain
C3 Imaging Research Center
University of Galway
Galway, Ireland
rishabh.jain@universityofgalway.ie

Peter Corcoran
C3 Imaging Research Center
University of Galway
Galway, Ireland
peter.corcoran@universityofgalway.ie

2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) | 979-8-3503-2797-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/SpeD59241.2023.10314899

***Abstract*— Speech synthesis technology has witnessed significant advancements in recent years, enabling the creation of natural and expressive synthetic speech. One area of particular interest is the generation of synthetic child speech, which presents unique challenges due to children's distinct vocal characteristics and developmental stages. This paper presents a novel approach that leverages the Fastpitch text-to-speech (TTS) model for generating high-quality synthetic child speech. This study uses the transfer learning training pipeline. The approach involved finetuning a multi-speaker TTS model to work with child speech. We use the 'cleaned' version of the publicly available MyST dataset (55 hours) for our finetuning experiments. We also release a prototype dataset of synthetic speech samples generated from this research together with model code to support further research. By using a pretrained MOSNet, we conducted an objective assessment that showed a significant correlation between real and synthetic child voices. Additionally, to validate the intelligibility of the generated speech, we employed an automatic speech recognition (ASR) model to compare the word error rates (WER) of real and synthetic child voices. The speaker similarity between the real and generated speech is also measured using a pretrained speaker encoder.**

Keywords—Fastpitch, synthetic speech, child speech, wav2vec2, MOSNet, Waveglow, MyST dataset.

I. INTRODUCTION

Speech synthesis technology has witnessed significant advancements in recent years, enabling the creation of natural and expressive synthetic speech. One area of particular interest is the generation of synthetic child speech, which presents unique challenges due to children's distinct vocal characteristics and developmental stages. Early research on Text-to-Speech (TTS) synthesis began several decades ago, primarily using concatenative and parametric methods [1]–[4]. While these methods generated speech from text, the resulting audio lacked naturalness and sounded robotic. Recent advancements in TTS models, mainly based on deep neural networks (DNN), have significantly improved the quality of synthesized speech. Tacotron [5], a neural sequence-to-sequence model, marked a notable improvement in speech synthesis quality. Subsequent models like Tacotron2 [6], FastSpeech [7], FastSpeech2 [8], FlowTTS [9], GlowTTS [10], Fastpitch [11], and Adaspeech [12] have further evolved TTS capabilities and improved TTS speech quality. Deepvoice2 [13] introduced the use of speaker verification models [14]–[16] to achieve multi-speaker TTS [17]–[21].

Child-TTS (CTTS), or TTS synthesis for child speech is currently limited due to the scarcity of child voice datasets and the challenges associated with their creation. Collecting child speech data for TTS research is challenging. Most TTS datasets are created in studios with expensive equipment, tailored for adult voices. While the pitch for adults typically falls between 70 to 250 Hz, children's speech ranges from 200 to 500 Hz [22]. Additionally, child speech exhibits distinct characteristics from adult speech, such as a higher fundamental frequency and variable speaking rates compared to adults [23]–[26]. Moreover, children tend to have longer phoneme durations and different prosody features due to their smaller vocal tracts [27]–[29].

This research aims to harness the potential of state-of-the-art (SOTA) TTS methods such as Fastpitch [11] to construct a pipeline for synthesizing children's voices while minimizing data requirements. The primary objective is to demonstrate the pipeline's ability to reliably generate a variety of self-consistent, distinct children's voices. Fastpitch utilizes a pitch prediction and duration prediction module which captures pitch variations in speech and enables more precise control over the speaking rate. This study uses an existing multispeaker children's speech dataset [30], which was cleaned to make it more suitable for CTTS research [31]. Subsequently, Fastpitch was trained on the cleaned dataset to generate synthetic speech for multiple child speakers, serving as a proof of concept.

By incorporating Fastpitch into the synthesis pipeline, we can effectively capture the unique prosodic features and intonation patterns present in speech. Our objective is to further optimize this model for child speech to accommodate individual characteristics, such as gender and regional accents, to produce realistic synthetic child voices. By using this approach, we intend to overcome the limitations of traditional TTS systems that often fail to capture the naturalness and authenticity of child-like speech. Our hypothesis is centered on the idea that pretraining the TTS model on adult speech data and subsequently finetuning it with child speech data can facilitate the synthesis of artificial child speech.

As part of this research, we also release a small set of synthetic datasets generated from this research. Objective evaluations were conducted on the synthesized child voices, comparing them to real child voices in terms of various acoustic features and Mean Opinion Score (MOS). The evaluation encompassed factors such as 'Naturalness', 'Intelligibility', and 'Speaker Similarity. Furthermore, we compared this approach with our previously reported Tacotron 2 TTS pipeline for the child speech synthesis [32]. In this study, no subjective evaluation was conducted; however, it will be taken into consideration for future research.

The potential applications of this research are wide-ranging and impactful such as educational tools, audiobooks

for children, language learning, interactive games and toys, virtual learning companions, and child-friendly voice assistants and chatbots to name a few. Such a pipeline would also enable the creation of large synthetic datasets, which could, in turn, enhance other areas of child speech research, such as speaker recognition and automatic speech recognition [33], [34].

II. Methodology

Fastpitch is a fully parallel TTS model conditioned on fundamental frequency contours. By incorporating Fastpitch into the synthesis pipeline, we can effectively capture the unique prosodic features and intonation patterns present in child speech. We present a multispeaker framework for TTS using a transfer learning approach that uses Waveglow vocoder for audio synthesis. We also evaluate this methodology using different objective evaluation methods to provide the validity of this approach. Fastpitch is used in this work due to its various advantages such as faster inference speed, improved prosody control, enhanced naturalness, duration control, multilingual support, and simplified architecture as compared to previous TTS approaches.

A. Datasets

In this section, we give an overview of the datasets used to finetune our pipeline and to implement some of our evaluation methods.

1) TTS Datasets:

These datasets are used for the TTS experiments for pretraining and finetuning the TTS model.

LibriTTS [35]: The LibriTTS corpus is an adult speech dataset that includes 585 hours of speech data sampled at a rate of 24kHz, obtained from a diverse set of 2,456 speakers. LibriTTS is widely used in research for training and evaluating text-to-speech systems.

MyST [30]: My Science Tutor (MyST) Corpus [36] is an American English child speech dataset from 1371 students containing over 393 hours of audio data out of which 197 hours are fully transcribed. We use the cleaned version of this dataset (derived from [31]), with 65 hours of speech divided into two subsets: 55 hours for training, called 'MyST_train' and 10 hours for testing, called 'MyST_test'. This 55 hours of training data is used for TTS training.

2) Text Datasets:

These datasets are used during inference as input text for the TTS model to generate data samples from the finetuned synthetic child voices.

Harvard Sentences [36]: Harvard sentences consist of 720 sentences that are carefully designed to be phonetically balanced. These sentences effectively encompass a wide range of phonemes.

LJ Speech Sentences [37]: This dataset contains 13,100 sentences extracted from the LJ Speech dataset.

B. Multispeaker child TTS using Fastpitch

1) Fastpitch (Acoustic Model) [11]:

FastPitch is a streamlined TTS model with a simplified encoder-decoder architecture, designed for faster inference and improved prosody control. In the multispeaker FastPitch TTS model, the input text is encoded using an encoder module, which typically comprises stacked layers of convolutional neural networks (CNNs) or recurrent neural networks (RNNs). The encoder processes the linguistic features of the text, such as phonemes or graphemes, and generates intermediate representations. The duration predictor module takes the intermediate representations from the encoder and predicts the duration of each phoneme or character in the input text. This enables the model to capture and generate natural speech rhythm and timing. The pitch predictor module takes the intermediate representations and predicts the fundamental frequency (F0) contour, controlling the pitch variations in the synthesized speech. The architecture of Fastpitch is detailed in Figure 1.

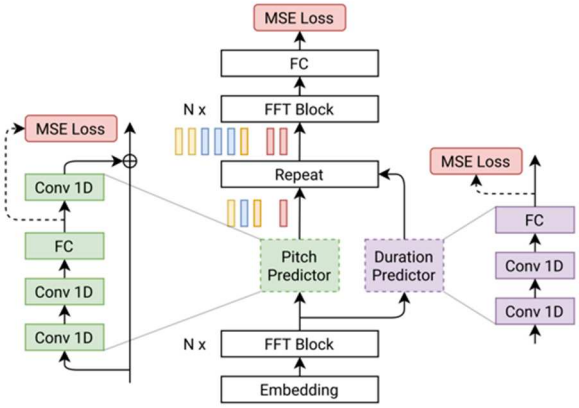


Fig. 1. Fastpitch Architecture [11].

We also condition the model on the speaker by adding a global speaker embedding [38] to the input tokens. The speaker embedding integration with the TTS framework [39]–[41] allows the model to capture the unique characteristics of different speakers. These embeddings encode speaker-related information in a vectorial representation for each speaker. During training, the model learns to associate speaker embeddings with the corresponding speakers, allowing it to generate speech that not only follows the desired linguistic content but also reflects the distinct vocal attributes of specific speakers. The primary loss function is the mean squared error (MSE) between the predicted mel-spectrogram and the target mel-spectrogram. Our work uses a newer version of Fastpitch, which is based on using the self-attention framework proposed in [38]. This enables the TTS model to learn speech-to-text alignment in parallel to TTS training instead of relying on an external aligner.

2) Transfer Learning Pipeline:

The proposed methodology involves pretraining the Fastpitch TTS model on a diverse dataset of adult speech, covering various age groups, linguistic backgrounds, and speech contexts. The LibriTTS dataset was used in this work. By finetuning the pretrained model on a smaller subset of the child speech dataset, such as MyST, will enable the model to learn the distinctive acoustic properties and pitch contours specific to child speech. Moreover, the model can be further optimized to accommodate individual characteristics, such as gender and regional accents, to synthesize more realistic CTTS voices.

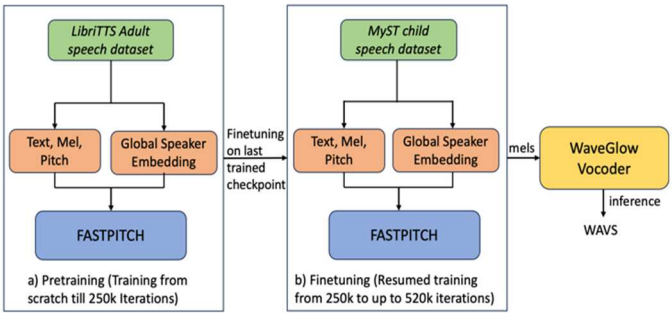


Fig. 2. Transfer learning pipeline: a) Pretraining: model being trained with LibriTTS dataset for up to 250k iterations. b) Finetuning: Resuming the acoustic model training with the MyST dataset from 250k iteration onwards up to 520k iteration.

The finetuning pipeline is kept consistent with our previous approach using Tacotron 2 [32] to allow for comparisons. Figure 2 describes the transfer learning pipeline. The model is first trained with the LibriTTS dataset (585 hours) for up to 250k iterations until a consistent low loss threshold is achieved, and the model starts to converge. After that, the model was finetuned for up to 520k additional steps using the MyST dataset (55 hours).

3) Waveglow (Vocoder) [42]:

WaveGlow is a SOTA vocoder model that generates high-quality and natural-sounding speech waveforms. It is based on a generative flow-based model architecture which models the distribution of speech waveforms. WaveGlow operates by taking a spectrogram representation of the speech as input and generating the corresponding waveform. The model employs an invertible neural network to transform the spectrogram into a latent space representation and then uses a series of invertible coupling layers to map this latent representation back into the waveform domain. Our WaveGlow model is trained on LibriTTS adult speech data which learns the complex relationship between spectrograms and waveforms. It was observed that Glow models [43]–[46] has popularly been used as a universal vocoder [45] and has been shown to work well with unseen speakers in multi-speaker models as well [47], [48]. Therefore, for the scope of this paper, WaveGlow (trained on LibriTTS) is used as a universal vocoder with synthetic child voices.

III. Experiments

A. Training details

The implementation is obtained from Nvidia's FastPitch Github[1]. For our training and finetuning process, we utilized two A6000 40GB GPUs. We employed a learning rate of 0.1 and a weight decay factor of 1e-6, maintaining consistency with their original implementation [11]. Additionally, the remaining hyperparameters were retained as per the provided implementation details. To ensure a smooth training process, we incorporated a warmup training step with a factor of 2000.

B. Experiments

1) Initial Experiments: These experiments involved using the LJ speech dataset for single-speaker finetuning. The model was first trained with LJ Speech and then finetuned with a single speaker from the MyST dataset. The output audio obtained was quite noisy. We also tried training the LJ speech single-speaker dataset and finetuning it with the complete MyST dataset (considering it as a single-speaker dataset). However, the results obtained didn't sound like child speech. Hence, finetuning on a single speaker was not explored further.

2) Main Experiments: These experiments involve multispeaker TTS training. The model was first trained with the LibriTTS dataset. Figure 3 shows an example loss curve of the LibriTTS training.

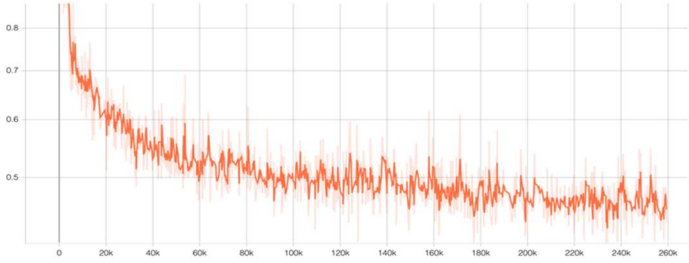


Fig. 3. LibriTTS pretraining curve (MSE loss vs. number of epochs)

It can be observed that for the first 2000 warmup steps, loss decreases gradually. After that loss decreases steadily until it reaches an average loss of 0.3 around 250k epoch. Since there was no improvement in loss function after that, it was decided to pause the training for further finetuning.

Further finetuning was performed from epoch 250k onwards on the MyST dataset. The loss increases until it starts to decrease around 260k epoch. From this point, there is a gradual decrease in loss until 520k steps. No significant improvement was observed in loss after this epoch. This was also verified by manually listening to generated audio files at an interval of every 50k epoch. After 550k epochs, the model exhibited signs of overfitting and began learning noise features in the MyST dataset, resulting in a decline in the quality of the synthesized audio.

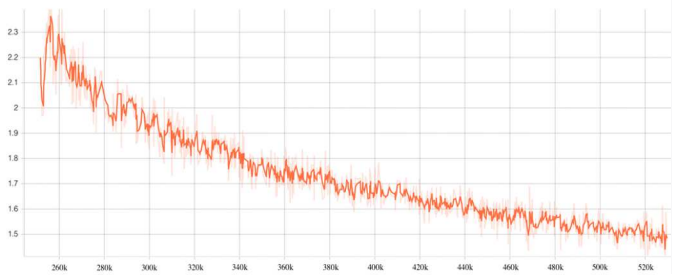


Fig. 4. MyST Finetuning curve (MSE loss vs. number of epochs).

C. Synthetic Datasets

We have generated two sets of synthetic child speech datasets. The dataset demographic is detailed in Table III. The dataset is made available through our GitHub[2]. Since the dataset was generated at a 22Khz sampling rate, FFmpeg was used to convert the data into a 16khz sampling rate for objective evaluation. The dataset is made available in both sampling rates. The dataset details are available below:

[1] https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/SpeechSynthesis/FastPitch

[2] https://github.com/C3Imaging/child_tts_fastpitch/

TABLE I. SYNTHETIC DATASET DEMOGRAPHICS

Dataset	*Speakers*	*Hours*	*Utterances*	*data/speaker*
CS_HS	40	29.02	28,800	43.53 minutes
CS_LJ	2	47.61	26,200	23.8 hours

1) CS_HS – This dataset used Harvard Sentences as a text reference to generate the synthetic child speech dataset. We selected the 40 speakers with the most amount of data in hours. from the LibriTTS dataset which was used to generate 40 child speakers. See Table 1 for more details.

2) CS_LJ – This dataset used LJ Speech transcripts as a text reference to generate the synthetic child speech dataset. We selected one male and one female speaker from the LibriTTS dataset which contained the most amount of training. These speakers were subjected to generate the child's speech. See Table 1 for more details.

IV. RESULTS AND EVALUATION

Our experimental findings demonstrate the successful synthesis of child voices using our proposed methodology. To assess the validity of the generated speech, we conducted objective evaluations, specifically focusing on the aspects of Naturalness, Intelligibility, and Speaker similarity. Furthermore, we conducted a comparative analysis with our previous research, which involves synthesizing child speech using the Tacotron 2 model. For the evaluation process, we randomly selected 120 utterances from the original MyST dataset, Tacotron-based synthetic dataset, and Fastpitch-generated synthetic utterances (from III.B). This allowed us to systematically compare the quality of speech generated by both the Tacotron 2 [32] and Fastpitch models within the context of child speech synthesis.

A. *Objective Naturalness Evaluation using the pretrained MOSNet [49]*

TABLE II. MOSNET OUTPUT FOR 120 SAMPLES WITH 95% CONFIDENCE INTERVAL

Dataset	*MOS*
Adult speech (Librispeech test_clean)	3.78 ± 0.07
Original Child Speech [MyST]	2.91 ± 0.07
Tacotron 2 based synthetic child speech [32]	2.60 ± 0.06
Fastpitch based synthetic child speech [Our work]	**3.10 ± 0.12**

Table 1 provides the Mean Opinion Scores (MOS) for 120 different speech samples using the pretrained MOSNet model [49]. MOSNet, trained on adult speech, exhibits a high correlation with human MOS ratings. However, its generalization to child speech is doubtful. Therefore, we only use MOSNet in this study to explore the correlation between reference child audio and synthetic child audio. It acts as a measure to validate the 'Naturalness' of the speech The original child speech from the MyST dataset received an average MOS of 2.91 ± 0.07, indicating moderate acceptability. The Fastpitch-generated child speech indicates higher quality than both the original speech and Tacotron2. These results suggest that the Fastpitch model, as implemented in our research, produces a strong correlation between synthetic child speech and real child speech.

B. *Objective Intelligibility Evaluation using a pretrained wav2vec2 ASR System [50]*

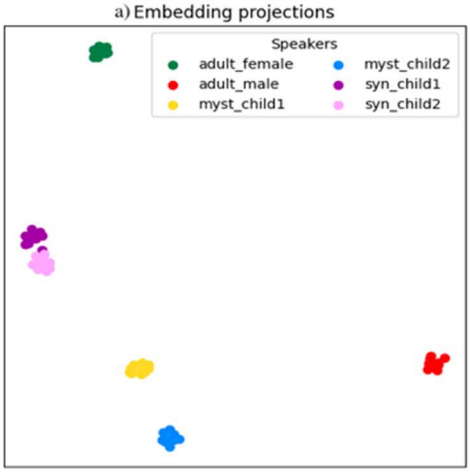
TABLE III. WER ON 120 RANDOMLY SELECTED UTTERANCES FROM ADULT SPEECH, REAL CHILD SPEECH, AND SYNTHETIC CHILD SPEECH USING THE WAV2VEC2 BASE ASR MODEL

Dataset	*WER*
Adult Speech (Librispeech test_clean)	3.43
Original Child Speech [MyST]	15.27
Tacotron 2 based synthetic child speech [32]	25.63
Fastpitch based synthetic child speech [Our work]	**17.61**

In this study, we employed the wav2vec2 base model [3], which was finetuned with 960 hours of the Librispeech dataset, to evaluate the 'Intelligibility' of the generated child speech. Since wav2vec2 is a SOTA ASR model, it was intended to use this as a validity metric for the synthetic speech. Additionally, we conducted a comparative analysis with our previous approach utilizing Tacotron 2 [32]. Table II provides WER for different speech datasets. The adult speech dataset achieved a strong WER of 3.43, considering the model's training on adult speech data. Our Fastpitch-based approach achieved a WER of 17.61, closely resembling the WER of the original child speech from the MyST dataset. Moreover, it surpassed the WER of the Tacotron 2 generated child speech, indicating improved performance over the synthetic child speech.

C. *Speaker similarity verification using a pretrained speaker verification system [15]*

Speaker similarity between a synthesized speech and a real speech can be calculated using a speaker verification system [15]. The pretrained speaker encoder from Resemblyzer[4] was used to extract and visualize the speaker embeddings. This tool uses cosine distance to calculate the similarity between the two embeddings.



[3] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

[4] https://github.com/resemble-ai/Resemblyzer

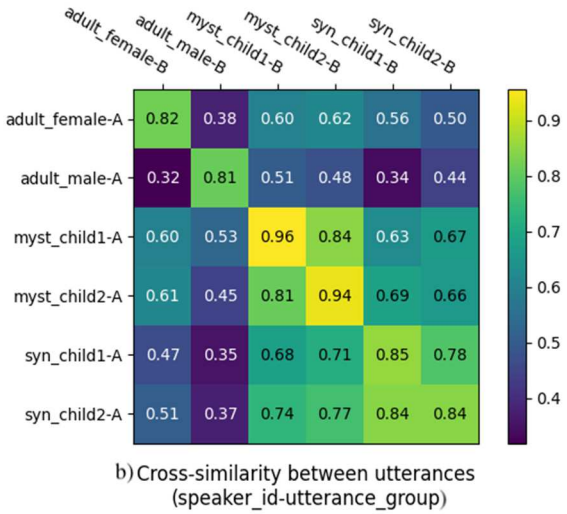


Fig. 5. a) Projections of embeddings between different real and synthetic child speech in comparison to adult speakers. b) Cross-similarity between 10 speakers in Set A and Set B.

For this evaluation, 6 speakers were randomly selected: 2 from LibriTTS [one male and one female], 2 from the MyST dataset, and 2 from the synthetically generated CS_HS dataset. We selected 10 utterances for each speaker in a random order. A visualization of this similarity in a 2D projection can be seen in Figure 5a. It can be observed that most of the child speakers (both real and synthetic) are very close in a cluster compared to adult male and female speakers.

To further demonstrate the similarity between real child speech and synthetic child speech, cosine similarity was used to calculate the cross-similarity between each speaker. All 6 speakers were divided into two sets, A and B. Speaker embeddings are extracted for each of the utterances for each of the sets and averaged together for each speaker. This gave us 6 unique speaker embeddings in sets A and B for each of the 6 speakers. Cosine similarity is finally used to measure the similarity between sets A and B. Figure 5b shows the plot for the cross similarity between 6 speakers. The similarity for most of the child and adult speech is between 0.34-0.53 whereas the similarity for synthetic child speech and real child speech is between 0.63-0.98. The average similarity between synthetic and real child voices is 77%. Hence, we can conclude that our synthetically generated child speech is quite close to real child speech in terms of speaker similarity.

V. Conclusion and Future Work

This paper presents a pipeline for synthesizing child speech in scenarios with limited training data. The proposed approach involves cleaning an existing child speech dataset to create a small, curated dataset suitable for TTS training. A transfer learning technique is employed, utilizing pretraining on adult speech data and finetuning on child speech data. Objective evaluations using MOSNet demonstrate a strong correlation between real and synthesized child voices. Using a pretrained adult speech wav2vec2 ASR model, the WER for synthetic child voices was measured at 17.61, compared to a WER of 15.27 for real child voices. Speaker similarity evaluation using a pretrained speaker encoder yields an average cosine similarity of 77% between synthetic speech and the original speakers. Synthetic child speech samples are available on the project's GitHub. We also release two small synthetic child speech datasets generated from this work. Multi-speaker TTS proves to be a valuable approach for child speech synthesis, even with limited training data.

For Future work, we aim to perform a subjective evaluation (as proposed in [32]) on the released dataset for better clarity over the 'Naturalness', 'Intelligibility', and 'Speaker Similarity' of the generated child speech. Furthermore, it is also intended to investigate the use of synthetically generated child speech to enhance other areas of child speech research, such as ASR and speaker recognition.

Acknowledgment

The authors would like to acknowledge experts from Xperi: Gabriel Costache, Zoran Fejzo, Francisco Salgado, and George Sterpu for providing their expertise and feedback throughout. The authors would also like to thank Adriana Stan and Horia Cucu from the University Politehnica of Bucharest, for providing her expertise on TTS/ASR experiments.

References

[1] O. Watts, J. Yamagishi, K. Berkling, and S. King, 'HMM-based synthesis of child speech', *Proc 1st Workshop Child Computer Interaction ICMI08 Post-Conf. Workshop*, 2008.

[2] O. Watts, J. Yamagishi, S. King, and K. Berkling, 'Synthesis of child speech with HMM adaptation and voice conversion', *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 1005–1016, 2010, doi: 10.1109/TASL.2009.2035029.

[3] A. W. Black, H. Zen, and K. Tokuda, 'Statistical parametric speech synthesis', in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, IEEE, 2007, p. IV-1229-IV–1232.

[4] Maia, R., Zen, H., Gales, M.J.F. (2010) Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters. Proc. 7th ISCA Workshop on Speech Synthesis (SSW 7), 88-93

[5] Y. Wang *et al.*, 'Tacotron: Towards End-To-End Speech Synthesis'.

[6] J. Shen *et al.*, 'Natural TTS Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions'.

[7] Y. Ren *et al.*, 'Fastspeech: Fast, robust and controllable text to speech', *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[8] Y. Ren *et al.*, 'FastSpeech 2: Fast and High-Quality End-to-End Text to Speech'. arXiv, Aug. 07, 2022.

[9] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, 'Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow', in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7209–7213. doi: 10.1109/ICASSP40776.2020.9054484.

[10] J. Kim, S. Kim, J. Kong, and S. Yoon, 'Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search'. Available: https://github.com/jaywalnut310/glow-tts.

[11] A. Łańcucki, 'Fastpitch: Parallel text-to-speech with pitch prediction', in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6588–6592.

[12] M. Chen *et al.*, 'Adaspeech: Adaptive text to speech for custom voice', *ArXiv Prepr. ArXiv210300993*, 2021.

[13] S. Arik, ''Deep voice 2: multi-speaker neural text-to-speech,'' in Proc. 31st Int. Conf. Neural Inf. Process. Syst. Red Hook, NY, USA, Curran Associates Inc., 2017, pp. 2966–2974.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, 'X-Vectors: Robust DNN Embeddings for Speaker Recognition', in *ICASSP, IEEE International Conference on Acoustics, Speech and*

Signal Processing - Proceedings, 2018, pp. 5329–5333. doi: 10.1109/ICASSP.2018.8461375.

[15] L. Wan Quan Wang Alan Papir Ignacio Lopez Moreno, 'Generalized End-To-End Loss for Speaker Verification' in ICASSP, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2018, pp. 4879–4883.

[16] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, ''End-to-end textdependent speaker verification,'' in ICASSP, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5115–5119.

[17] M. Chen *et al.*, 'Cross-lingual, Multi-speaker Text-To-Speech Synthesis Using Neural Speaker Embedding', 2019, doi: 10.21437/Interspeech.2019-1632.

[18] R. Valle, J. Li, R. Prenger, and B. Catanzaro, ''Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens,'' in ICASSP, *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2020, pp. 6189–6193.

[19] Y. Jia *et al.*, 'Transfer learning from speaker verification to multispeaker text-to-speech synthesis', in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2018, pp. 4480–4490.

[20] A. Kulkarni, V. Colotte, and D. Jouvet, ''Improving latent representation for end-to-end multispeaker expressive text to speech system,'' Tech. Rep. ffhal-02978485v1f, 2020.

[21] E. Cooper, C.-I. Lai, Y. Yasuda, and J. Yamagishi, ''Can speaker augmentation improve multi-speaker end-to-end TTS?'' in Proc. Interspeech, Oct. 2020, pp. 1–5.

[22] S. Shahnawazuddin, N. Adiga, and H. K. Kathania, 'Effect of Prosody Modification on Children's ASR', *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1749–1753, Nov. 2017, doi: 10.1109/LSP.2017.2756347.

[23] G. Yeung, R. Fan, and A. Alwan, ''Fundamental frequency feature normalization and data augmentation for child speech recognition,'' in ICASSP, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, 2021, pp. 6993–6997, doi: 10.1109/ICASSP39728.2021.9413801.

[24] S. Lee, A. Potamianos, and S. Narayanan, ''Acoustics of children's speech: Developmental changes of temporal and spectral parameters,'' J. Acoust. Soc. Amer., vol. 105, no. 3, pp. 1455–1468.

[25] S. Shahnawazuddin, R. Sinha, and G. Pradhan, 'Pitch-Normalized Acoustic Features for Robust Children's Speech Recognition', *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1128–1132, Aug. 2017, doi: 10.1109/LSP.2017.2705085.

[26] S. Lee, A. Potamianos, and S. S. Narayanan, 'Analysis of children's speech: duration, pitch and formants', in *EUROSPEECH*, 1997.

[27] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, ''A review of ASR technologies for children's speech,'' in Proc. 2nd Workshop Child, Comput. Interact. (WOCCI), 2009, pp. 1–8.

[28] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, ''Acoustic normalization of children's speech,'' in Proc. 8th Eur. Conf. Speech Commun. Technol., 2003.

[29] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, ''Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition,'' in Proc. IEEE Int. Conf. Acoustics Speech Signal Process., 2006, pp. I–I, doi: 10.1109/ICASSP.2006.1660040.

[30] W. Ward, R. Cole, and S. Pradhan, ''My science tutor and the MyST corpus,'' Tech. Rep., 2019.

[31] R. Jain, A. Barcovschi, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, 'A Wav2vec2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition', Apr. 2022, doi: 10.48550/arxiv.2204.05419.

[32] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, 'A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis', *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.

[33] K. Yang, T.-Y. Hu, J.-H. R. Chang, H. Swetha Koppula, and O. Tuzel, 'Text is all You Need: Personalizing ASR Models Using Controllable Speech Synthesis', in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096971.

[34] A. Fazel *et al.*, 'SynthASR: Unlocking Synthetic Data for Speech Recognition', in *Interspeech 2021*, ISCA, Aug. 2021, pp. 896–900. doi: 10.21437/Interspeech.2021-1882.

[35] H. Zen *et al.*, 'LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech'. arXiv, Apr. 05, 2019.

[36] 'IEEE Recommended Practice for Speech Quality Measurements', *IEEE Trans. Audio Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969, doi: 10.1109/TAU.1969.1162058.

[37] 'The LJ Speech Dataset'. https://keithito.com/LJ-Speech-Dataset.

[38] R. Badlani, A. \Lańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, 'One TTS alignment to rule them all', in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6092–6096.

[39] P. Neekhara, J. Li, and B. Ginsburg, 'Adapting TTS models For New Speakers using Transfer Learning'. arXiv, Apr. 05, 2022. Available: http://arxiv.org/abs/2110.05798

[40] F. Lux, J. Koch, and N. T. Vu, 'Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech', in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 962–969. doi: 10.1109/SLT54892.2023.10022433.

[41] C.-P. Hsieh, S. Ghosh, and B. Ginsburg, 'Adapter-Based Extension of Multi-Speaker Text-to-Speech Model for New Speakers'. arXiv, Nov. 01, 2022.Available: http://arxiv.org/abs/2211.00585

[42] R. Prenger, R. Valle, and B. Catanzaro, ''Waveglow: A flow-based generative network for speech synthesis,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 3617–3621.

[43] W. Jang, D. Lim, and J. Yoon, 'Universal melgan: A robust neural vocoder for high-fidelity waveform generation in multiple domains', *ArXiv Prepr. ArXiv201109631*, 2020.

[44] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, 'Universal Neural Vocoding with Parallel Wavenet', in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 6044–6048. doi: 10.1109/ICASSP39728.2021.9414444.

[45] J. Lorenzo-Trueba *et al.*, 'Towards achieving robust universal neural vocoding', *ArXiv Prepr. ArXiv181106292*, 2018.

[46] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, 'Universal neural vocoding with parallel wavenet', in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6044–6048.

[47] D. Paul, Y. Pantazis, and Y. Stylianou, 'Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions', Accessed: Mar. 10, 2022. [Online]. Available: https://github.com/fatchord/WaveRNN

[48] P. L. Tobing and T. Toda, 'High-fidelity and low-latency universal neural vocoder based on multiband WaveRNN with data-driven linear prediction for discrete waveform modeling', 2021.

[49] C.-C. Lo *et al.*, 'MOSNet: Deep Learning based Objective Assessment for Voice Conversion', in *Interspeech 2019*, Sep. 2019, pp. 1541–1545. doi: 10.21437/Interspeech.2019-2003.

[50] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, 'wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 12449–12460.

Appendix E

Exploring Native and Non-Native English Child Speech Recognition with Whisper.

Authors: Rishabh Jain (RJ), Andrei Barcovschi (AB), Mariam Yiwere (MY), Peter Corcoran (PC) and Horia Cucu (HC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	RJ: 80%, AB:10%, PC:10%
Experiments and Implementation	RJ: 90% AB:10%
Background	RJ: 90%, AB:10%
Manuscript Preparation	RJ: 70%, AB: 10%, MY: 5%, PC: 10%, HC: 5%

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Exploring Native and Non-Native English Child Speech Recognition with Whisper

Rishabh Jain[1], (Graduate Student Member, IEEE), Andrei Barcovschi[1], Mariam Yiwere[1], Peter Corcoran[1], (Fellow, IEEE) and Horia Cucu[2], (Member, IEEE)

[1] School of Electrical and Electronics Engineering, University of Galway, Galway, H91 TK33 Ireland
[2] Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Romania

Corresponding author: Rishabh Jain (e-mail: rishabh.jain@universityofgalway.ie).

This work was supported by the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project (2020–2023) funded by the Disruptive Technologies Innovation Fund (DTIF), the College of Science & Engineering Ph.D. Research Scholarship at the University of Galway, and the SFI ADAPT Center for Digital Media Research (Grant No. 13/RC/2106_P2).

ABSTRACT Modern end-to-end Automatic Speech Recognition (ASR) systems struggle to recognise children's speech. This challenge is due to the high acoustic variability in children's voices and the scarcity of child speech training data, particularly for accented or low-resource languages. This study focuses on improving the performance of ASR on native and non-native English child speech using publicly available datasets. We evaluate how the large-scale whisper models (trained with a large amount of adult speech data) perform with child speech. In addition, we perform finetuning experiments using different child speech datasets to investigate the performance of whisper ASR on non-native English-speaking children's speech. Our findings indicate relative Word Error Rate (WER) improvements ranging from 29% to 89% over previous benchmarks on the same datasets. Notably, these gains were achieved by finetuning with only a 10% sample of unseen non-native datasets. These results demonstrate the potential of whisper for improving ASR in a low-resource scenario for non-native child speech.

INDEX TERMS child automatic speech recognition, whisper, large-scale supervision, MyST, PFSTAR, CMU_Kids, speechocean762, non-native child speech.

I. INTRODUCTION

While ASR performance for adult speech has improved in recent times due to the availability of large-scale transcribed speech corpora and the development of end-to-end (E2E) attention-based acoustic models [1]–[4], the same benefits have not been extended to the child speech domain due to a lack of available transcribed child audio data. The acoustic variability in children's speech caused by developmental changes of the vocal tract coupled with the child's limited linguistic and phonetic knowledge affects the performance of ASR systems for this age group [5]–[8]. Furthermore, the scarcity of data for ASR training in the child speech domain is an acute problem, as acquiring and annotating such data is a complex and resource-intensive task [9].

Recent developments in transfer learning have shown promising results in ASR, especially in recognizing speech from low-resource languages [10]–[12]. A key strategy involves finetuning an acoustic model. The model leverages frame-level acoustic representations derived from self-supervised models like wav2vec2 [3], [13], which were initially trained on vast amounts of unlabeled adult speech data using a masking objective. This has proven to be effective for downstream speech recognition applications with small amounts of labelled data. However, the self-supervised learning (SSL) training procedure for ASR is less effective in the case of domain shifting [14]. This means that the performance drops when the model encounters data that significantly differs from the training set, such as non-native child speech, making accurate recognition difficult.

Supervised transfer learning has also emerged as a promising solution to this problem. It adapts features learned from adult speech to enhance child speech recognition [15]–[18]. Additionally, audio augmentation techniques, which expand the training dataset [19]–[21], have also been effective in boosting ASR performance for child speech. Recent work on ASR for non-native child speech has also explored transfer learning as a way to make significant improvements [15], [19]–[21]. For instance, the

use of a pretrained transformer model for transfer learning has been investigated to better adapt to non-native children's speech [18]. Moreover, there have been notable strides in supervised learning approaches that show potential for child speech recognition [16], [22], [23]. Previous studies [24], [25] have demonstrated that training models across multiple datasets using supervised learning methods can enhance the model's ability to generalize across new, unseen datasets. This broad approach to training suggests a pathway towards more robust and adaptable ASR systems capable of handling the complexities of unseen child speech.

Given the low-resource nature of child speech and the limited datasets available for research use, this study opted to utilize the recent state-of-the-art (SOTA) Whisper [4] approach. Whisper has successfully addressed the challenges of weakly supervised speech recognition by training on large amounts of labelled adult audio datasets in a supervised manner. It has shown impressive performance in low-resource multilingual languages due to its multitask learning objectives and the use of multilingual datasets for training [4]. This research aims to investigate whether Whisper's multilingual training approach can enhance ASR performance in the particularly challenging area of low-resource child speech. First, we evaluate the performance of the original pretrained whisper models on different native and non-native English child speech datasets. Since whisper learns speech representations from a large number of multilingual audio datasets, it was also intended to adapt these whisper models for non-native English child speech datasets by performing further finetuning.

The primary contribution of this paper lies in adapting and finetuning the whisper model for child speech recognition. While finetuning large transformer models on small datasets is a well-established practice, our study goes beyond this by focusing on the unique characteristics and challenges associated with non-native child speech data. Through the careful crafting of experiments, we demonstrate the effectiveness of the whisper transformer model and wish to underscore the practical implications of our research, such as the promising applications of the finetuned model in real-world scenarios. Child speech recognition has wide-ranging applications in the education, healthcare, and accessibility domains. The main contributions of the paper are highlighted as the following:

- Demonstrates significant performance improvements on non-native English child speech datasets.
- Showcases whisper's ability to adapt effectively to diverse child speech datasets through finetuning.
- Proves whisper's resistance to catastrophic forgetting, maintaining performance on adult speech while improving child speech recognition.
- Provides insightful analysis and discussions of the outcomes derived from whisper finetuning.

II. METHODOLOGY

In this work, the whisper [4] model is used, showcasing the benefits of large-scale weakly supervised pre-training for improved ASR performance. It employs training data of up to 680,000 hours of labelled audio data, of which 117,000 hours include 96 non-English languages and 125,000 hours of X→en translation data.

A. Whisper Architecture

The architecture of the whisper model (see Figure 1) is based on an encoder-decoder transformer, which uses 80-channel log-Mel spectrograms as input. The encoder consists of two convolution layers with a kernel size of 3, a sinusoidal positional encoding, and a stacked set of transformer blocks. The decoder also uses the learned positional embeddings and the same number of transformer blocks as the encoder. The model uses a byte-level BPE text tokenizer [26] for English-only models and refits the vocabulary for multilingual models to avoid excessive fragmentation in other languages. A multitask training format is used, where models are trained to perform various speech-processing tasks using a single decoder. Multitask training is done by conditioning the decoder on a sequence of input tokens that specify the task and desired output format. These tasks include multilingual speech recognition, spoken language detection, speech translation, and voice activity detection.

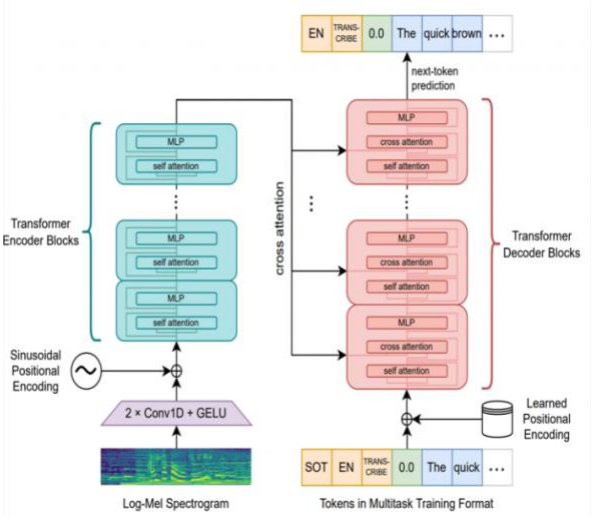


FIGURE 1. Whisper: Sequence-to-sequence Transformer model trained on multitask learning (figure from [4]).

B. Training Details

The models were optimized using AdamW [27] and gradient norm clipping [28] with a linear learning rate decay with a warmup over the first 2048 updates. The pretrained whisper models are categorized based on their sizes, namely: tiny, base, small, medium, and large (see Table 1). There are two versions of each model: one trained with multilingual data and one using only English data (indicated by '.en' in the name). We provide the initial non-finetuning results on all

the available pretrained models. We select the best-performing models and apply finetuning to those. Architectural hyperparameter details can be found in Table 1.

TABLE 1
ARCHITECTURE PARAMETERS FOR WHISPER

Models	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	72M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Note: After the release of the initial whisper models, the authors trained the 'large' model for an additional 2.5X epochs, incorporating various regularization techniques [4]. This updated model is denoted as Large-V2.

Whisper is trained using a large amount of multilingual speech datasets including low-resource languages, and we aimed to investigate whether this multilingual-focused ASR model could be used to improve performance on non-native child speech. We performed finetuning using parallel child audio data on the final layer of the pretrained whisper models for up to 4000 epochs [4], with a learning rate of 1e-05 and a linear learning rate scheduler.

C. Decoding in Whisper

The whisper ASR system employs several decoding strategies during inference [4], and these strategies are executed up to six times. The goal is to select the best transcription based on the heuristics and the decoding strategies' performance.

1. Beam Search with 5 Beams: This strategy uses beam search, a common technique in ASR systems. It explores multiple hypotheses (in this case, five) and selects the one with the highest log probability as the final transcription. This approach favours more probable sequences.
2. Greedy Decoding with Best of 5 Sampling: Greedy decoding starts with the most likely token at each step, while sampling introduces randomness. The system uses a sampling temperature schedule (0.0, 0.2, 0.4, 0.6, 0.8, 1.0) for successive attempts. Lower temperatures make the sampling approach more deterministic, while higher temperatures allow more randomness in token selection. This strategy explores a range of sampling behaviours to find the most suitable transcription.

These decoding strategies are applied to enhance transcription quality, particularly in situations where the model may be less certain, such as when there is background noise or other challenging audio conditions are present. The impact of these decoding strategies can vary across different datasets, as noted in the whisper paper [4], but collectively, they help improve transcription accuracy and reliability by considering both the model's confidence and the compression characteristics of the transcribed text. We do not use any external language models since it was intended to keep the decoding technique identical to the original implementation by the whisper authors [4] and concentrate on the recognition capabilities of the finetuned acoustic models.

III. CORPUS DESCRIPTION

The authors of Whisper do not explicitly mention the list of training datasets used for pretraining [4]. We used the following child speech datasets for our finetuning and testing experiments: MyST Corpus [29], PFSTAR Corpus [30], CMU_Kids Corpus [31], and Speechocean672 [32]. LibriTTS [33] was the only adult dataset used in the experiments during inference.

A. Dataset Cleaning and Description

Each dataset was cleaned according to whisper authors' text standardization guidelines [4]. The abbreviations, punctuations, white spaces, non-linguistic symbols, and other non-alphanumeric characters were removed from the transcripts, and all the characters were changed to lowercase. All the audio data was converted to a 16-bit mono channel with a 16Khz sampling rate and saved as '.wav' audio files, while the transcriptions were saved as '.txt' files. Child data-specific cleaning methodology was kept consistent with [34]. Given the low resource nature of non-native child speech datasets, we opted to split the available data into 80% for testing and 20% for training. Allocating a larger proportion of the data for testing helped obtain more objective results. The datasets used are described below:

1) LibriTTS [33] is a multispeaker English adult speech dataset. The 'dev-clean' subset of LibriTTS with 9 hours of audio is used as the representative for adult speech for our finetuned models during testing.

2) My Science Tutor (MyST) Corpus [29] is an American English child speech dataset containing over 393 hours of audio data out of which 197 hours are fully transcribed. We use the cleaned version of this dataset (as described in [34]), with 65 hours of speech divided into two subsets: 55 hours for training, called 'MyST_train' and 10 hours for testing, called 'MyST_test'.

3) PFSTAR Corpus [30] contains a collection of words spoken by native British English children and non-native English child speech from Swedish, German, and Italian natives. The cleaned PFSTAR British dataset (as described in [34]) contains a total of 12 hours of usable audio. This data was divided into 10 hours for training, called 'PF_br_train', and 2 hours for testing, called 'PF_br_test'.

The PFSTAR Swedish subset contains 1.27 hours of English child speech with Swedish accents. It is divided into 1.01 hours (80%) for testing and 0.24 hours (20%) for training. Testing and training subsets are named 'PF_sw_test' and 'PF_sw_train', respectively. The

PFSTAR German subset contains 3.4 hours of English child speech with German accents and is divided into 2.55 hours (80%) for testing and 0.68 hours (20%) for training, referred to as 'PF_ge_test and 'PF_ge_train', respectively. The PFSTAR Italian subset, containing 3.5 hours of Italian-accented English child speech, is divided into 2.8 hours (80%) for testing and 0.7 hours (20%) for training, called 'PF_it_test' and 'PF_it_train', respectively.

4) CMU_Kids [31] contains 9 hours of read-aloud sentences recorded by children at Carnegie Mellon University. This was divided into 7 hours for training, called 'CMU_train' and 2 hours for testing, called 'CMU_test'. Compared to other child speech datasets, the CMU_Kids dataset is quite noisy and contains many different types of background noises.

5) Speechocean762 [32] contains non-native English speech from Chinese-accented speakers of different age groups. We selected speakers whose age was less than 18, amounting to 2.4 hours of speech divided into 1.92 hours (80%) for testing and 0.48 hours (20%) for training, named 'SO_test' and 'SO_train' respectively.

B. Dataset Usage in Training and Testing

MyST, PFSTAR British, and CMU_Kids are considered native English child speech datasets, while PFSTAR (Swedish, German, and Italian) and Speechocean762 are non-native English child speech datasets in this study. The dataset division into training and testing categories is presented in Table 2.

TABLE 2
DATASETS USED FOR TRAINING AND TESTING

Finetuning/Training Datasets	Test Datasets
MyST_train (55 hours)	MyST_test (10 hours)
PF_br_train (10 hours)	PF_br_test (2 hours)
CMU_train (7 hours)	CMU_test (2 hours)
PF_sw_train (0.24 hours)	PF_sw_test (1.01 hours)
PF_ge_train (0.68 hours)	PF_ge_test (2.55 hours)
PF_it_train (0.7 hours)	PF_it_test (2.8 hours)
SO_train (0.48 hours)	SO_test (1.92 hours)
NN_50 (1.06 hours)	dev-clean (9 hours)
NN_100 (2.13 hours)	

Due to the limited volume of non-native data, we consolidated the non-native training datasets described earlier into two distinct subsets for the purpose of finetuning:

Non_Native_10 (NN_10): This subset comprises half of the selected non-native training sets, specifically PF_sw_train, PF_ge_train, PF_it_train, and SO_train. It represents 10% of the overall non-native data pool.

Non_Native_20 (NN_20): This subset encompasses the entire range of non-native training datasets mentioned, including PF_sw_train, PF_ge_train, PF_it_train, and SO_train in their entirety. This constitutes 20% of the total non-native dataset.

IV. CODEBASE AND EXPERIMENTS

A. CODEBASE

The whisper finetuning codebase, used for implementing our initial testing and subsequent finetuning is available here[1]. Our trained whisper models are openly available to use on the Hugging Face platform[2]. The information regarding the checkpoint, model parameters, learning rates, training curves, dataset availability, and access to cleaned datasets are available on our GitHub[3]. We followed the same finetuning approach as in our earlier work with the whisper model [35]. This study is essentially a continuation of our previous research, where we specialized whisper models for recognizing children's speech and compared it with the wav2vec2 self-supervised approach on the same distribution of datasets.

Nine sets of experiments were conducted, organized into groups A, B, C, D, E, F, G, H and I as detailed below. Table 3 shows the Word Error Rates (WERs) obtained from these experiments, a standard metric for evaluating ASR system performance. WER quantifies the error rate in recognizing spoken words against a reference transcript, calculated by summing substitution, deletion, and insertion errors, then dividing by the total word count in the reference. A lower WER indicates better performance.

Group A served as the baseline, comprising tests on the original Whisper models without finetuning. This establishes the benchmark performance. The remaining groups (B, C, D, E, F, G, H, and I) focused on the three top-performing models from group A, which were finetuned using different distributions of child speech training datasets as detailed in Table 2. Various experiments were conducted to finetune the ASR models by utilizing different combinations of child speech datasets. The objective was to identify the optimal combinations of child audio training data that would result in the lowest WERs on diverse test datasets. Additionally, different data distributions were employed to determine the complementary datasets and identify those that hindered the improvement of the ASR model.

In Group B, the models were subjected to finetuning using the MyST_train dataset, which was selected for the initial finetuning experiments as it is the largest available child speech dataset. Subsequently, in Group C and Group D experiments, the next largest datasets, namely CMU_train and PF_br_train, were added along with MyST_train. Group C models used the MyST_train and CMU_train

[1] **Whisper Implementation:** https://github.com/huggingface/community-events/tree/main/whisper-fine-tuning-event
[2] **Finetuned Whisper models:** https://huggingface.co/rishabhjain16
[3] **GitHub**: https://github.com/C3Imaging/whisper_non_native_child_asr

datasets, while Group D models used the MyST_train and PF_br_train combination of datasets. For Group E finetuning experiments, all three datasets, namely MyST_train, CMU_train, and PF_br_train, were used collectively. The remaining experiments focused on finetuning using different distributions of Non-Native English child speech datasets (NN_10 and NN_20), to study the performance of the finetuned models on the test datasets. Thus, in Group F, the models were finetuned using the MyST_train, PF_br_train, and NN_10 datasets, while in Group G, the models were finetuned using the MyST_train, PF_br_train, and NN_20 datasets. Finally, we wanted to assess the ASR performance using the complete set of available datasets, therefore, Group H models were finetuned using the MyST_train, CMU_train, PF_br_train, and NN_10 datasets. Similarly, Group I models were finetuned with the MyST_train, CMU_train, PF_br_train, and NN_20 datasets.

V. RESULTS AND DISCUSSIONS

A. MAIN RESULTS FROM GROUP EXPERIMENTS

The results obtained from these experiments are presented in Table 3 and the lowest WERs are highlighted in bold.

1) GROUP A

The results from Group A highlight the WERs of pre-trained Whisper models across various speech datasets as outlined in Table 2. The findings show that smaller models, including Tiny, Base, and Small, generally exhibit higher WERs compared to the larger models, namely Medium and Large, as documented in Table 3. This trend suggests that the larger models, due to their increased size, possess a greater capacity for generalization, thereby enhancing speech recognition accuracy. When comparing models of equivalent size, it was observed that English-only models outperform their multilingual counterparts. This indicates that models trained specifically on language-focused datasets exhibit improved performance for those particular languages. Based on these insights, the models demonstrating the most robust performance — specifically the 'Medium', 'Medium.en', and 'Large-V2' models — were chosen for subsequent finetuning experiments.

2) GROUP B

The finetuning of models with the MyST_train dataset in Group B resulted in notable enhancements in ASR performance across all test datasets, with the sole exception of the CMU_test dataset. This could indicate a mismatch between the characteristics of the CMU_test data and the training data used for finetuning, possibly due to accent, dialect, or speech complexity differences not adequately covered by the MyST_train dataset. The ‘medium’ model showed a marked reduction in WER across various datasets, including a significant drop from baseline figures in Group A, highlighting the model's improved adaptability to different speech patterns post-finetuning. Overall, the average performance across Group B models illustrates the tangible benefits of finetuning on child speech, with improvements evident in lower WERs for a majority of the test scenarios.

3) GROUP C

In Group C, the finetuning process incorporated the CMU_train dataset alongside MyST_train, aiming to investigate its impact on reducing the WER for CMU_test. Following this finetuning, there was a notable decrease in the WER for the CMU_test dataset to as low as 2.32. However, this adjustment resulted in increased WERs across all other test datasets. Interestingly, the performance on the dev-clean dataset, which represents adult speech, remained unchanged. These outcomes hint at the acoustic similarities between the CMU_Kids dataset and adult speech, evidenced by the stable performance on adult speech and increased WERs on child speech datasets. This also suggests that the CMU_Kids dataset, while beneficial for targeting specificities of the CMU_test, may not align well with the acoustic properties of other child speech datasets. The disparity in WERs could also be linked to the inherent differences in domain or unique acoustic features present in the non-native test datasets, which are not sufficiently represented in the CMU_train dataset. Furthermore, the presence of low-quality audio within the CMU_train dataset might have adversely affected the model's performance, particularly evident in the heightened WERs observed in the PF_sw_test, PF_ge_test, PF_it_test, and SO_test datasets. This indicates that while targeted finetuning can enhance performance on specific datasets, it also underscores the challenge of balancing improvements across diverse speech datasets, especially when dealing with varying audio quality and distinct acoustic characteristics.

4) GROUP D

The decision to incorporate the PF_br_train dataset, a British English child speech dataset, into the finetuning process for Group D was influenced by the observed increase in WER across nearly all test datasets (except for the CMU_test), following the inclusion of CMU_train in Group C. This shift also aimed to assess the impact of PF_br_train on model performance across various speech recognition tasks. The results from Group D finetuning demonstrate a marked improvement across all non-native child speech test datasets, with WER decreasing for all tests except for the CMU_test. This suggests that the PF_br_train dataset's characteristics are more aligned with the acoustic properties required for effective recognition of non-native child speech, enhancing the models' performance significantly. The lower WERs in Group D can be attributed to the complementary nature of the PF_br_train dataset.

TABLE 3
WER FOR WHISPER ORIGINAL AND FINETUNED MODELS OVER DIFFERENT CHILD SPEECH TEST DATASETS USED IN THIS PAPER

ID	Models	MyST_test	PF_br_test	CMU_test	PF_sw_test	PF_ge_test	PF_it_test	SO_test	Dev_clean
Group A: No-Finetuning:									
1	Tiny	40.09	159.57	24.62	55.32	103.68	70.57	64.83	10.85
2	Tiny.en	33.02	47.11	16.25	45.23	89.80	47.22	51.28	8.62
3	Base	32.14	100.07	16.65	53.88	126.84	50.29	60.39	8.14
4	Base.en	29.15	45.70	15.01	37.29	93.77	46.84	38.47	7.18
5	Small	26.22	111.75	9.30	60.81	86.72	44.09	36.19	6.43
6	Small.en	26.72	39.00	8.64	32.26	**71.04**	33.38	30.33	6.06
7	**Medium**	**25.11**	80.97	**7.48**	35.07	105.82	45.65	37.00	**5.58**
8	**Medium.en**	28.06	**35.25**	**7.17**	**27.91**	80.40	**25.94**	**25.29**	6.20
9	Large	25.24	84.52	7.56	33.09	79.14	51.82	37.25	5.53
10	**Large-V2**	**25.00**	73.68	**6.86**	**29.99**	**77.56**	**34.97**	**29.39**	**5.40**
Group B: MyST_train Finetuning:									
11	Medium	**11.66**	19.76	9.43	34.18	**62.40**	24.53	24.89	5.62
12	Medium.en	11.81	17.83	**9.13**	**23.63**	76.84	**19.99**	25.45	6.48
13	Large-V2	12.28	**10.88**	9.80	25.56	65.58	23.48	**25.05**	**4.82**
Average (Group-B)		*11.91*	*16.15*	*9.45*	*27.79*	*68.27*	*22.67*	*25.13*	*5.64*
Group C: MyST_train + CMU_train Finetuning:									
14	Medium	12.14	41.83	4.46	158.75	113.07	125.05	33.24	6.10
15	Medium.en	**12.10**	31.29	**2.27**	138.95	125.37	77.38	33.32	6.13
16	Large-V2	12.37	**23.62**	2.32	184.24	211.01	180.79	48.34	**4.81**
Average (Group-C)		*12.20*	*32.24*	*3.01*	*160.64*	*149.81*	*127.74*	*38.3*	*5.68*
Group D: MyST_train + PF_br_train Finetuning:									
17	Medium	**12.22**	**2.98**	16.05	**16.52**	**51.53**	14.08	**22.80**	5.40
18	Medium.en	12.33	3.32	**15.08**	17.48	59.94	**13.95**	23.41	**4.88**
19	Large-V2	13.34	4.17	17.11	26.55	58.37	20.24	24.94	4.97
Average (Group-D)		*12.63*	*3.49*	*16.08*	*20.18*	*56.61*	*16.09*	*23.71*	*5.08*
Group E: MyST_train + CMU_train + PF_br_train Finetuning:									
20	Medium	11.72	3.11	2.36	23.94	86.13	16.72	27.88	5.62
21	Medium.en	**11.71**	**3.02**	2.23	**21.65**	**68.10**	**15.87**	**26.43**	5.57
22	Large-V2	12.37	3.10	**1.86**	43.34	71.18	56.29	32.99	**4.75**
Average (Group-E)		*11.93*	*3.07*	*2.15*	*29.64*	*75.13*	*29.62*	*29.10*	*5.31*
Group F: MyST_train + PF_br_train + NN_10 Finetuning:									
23	Medium	**11.73**	**3.15**	**9.33**	9.12	34.59	**5.10**	**16.02**	5.33
24	Medium.en	11.81	3.36	9.58	10.37	35.27	6.22	17.04	4.95
25	Large-V2	12.75	7.05	9.71	**8.39**	**33.48**	5.63	16.67	**5.09**
Average (Group-F)		*12.09*	*4.52*	*9.54*	*9.29*	*34.44*	*5.65*	*16.57*	*5.12*
Group G: MyST_train + PF_br_train + NN_20 Finetuning:									
26	Medium	11.96	3.12	**8.92**	7.74	36.21	4.16	14.40	5.39
27	Medium.en	12.30	3.28	9.53	8.94	34.78	4.42	14.87	5.01
28	Large-V2	**11.60**	**3.09**	9.22	**7.24**	**31.46**	**3.98**	**13.83**	**4.47**
Average (Group-G)		*11.95*	*3.16*	*9.22*	*7.97*	*34.15*	*4.18*	*14.36*	*4.95*
Group H: MyST_train + CMU_train + PF_br_train + NN_10 Finetuning:									
29	Medium	12.75	**3.11**	**1.98**	**8.99**	36.67	**5.14**	16.09	6.09
30	Medium.en	12.35	3.42	2.06	9.04	35.92	5.84	17.55	5.28
31	Large-V2	**11.73**	3.13	2.56	9.67	**35.05**	5.51	**15.83**	**4.69**
Average (Group-H)		*12.27*	*3.22*	*2.20*	*9.23*	*35.88*	*5.50*	*16.49*	*5.35*
Group I: MyST_train + CMU_train + PF_br_train + NN_20 Finetuning:									
31	Medium	12.55	3.09	1.96	**7.66**	34.77	**4.11**	**14.31**	6.06
33	Medium.en	11.88	3.28	1.98	8.16	34.99	4.65	15.87	5.15
34	Large-V2	**11.62**	**2.84**	**1.75**	8.36	**34.26**	4.40	14.52	**4.53**
Average (Group-I)		*12.01*	*3.07*	*1.89*	*8.06*	*34.67*	*4.38*	*14.9*	*5.25*

5) GROUP E

In the Group E experiments, we used a combination of MyST_train, CMU_train, and PF_br_train. In comparison to the results from Groups C and D, there is a performance increase on all the seen datasets, however, performance degradation can be observed on non-native datasets. This confirms that CMU_train datasets had a negative impact on the performance of non-native English test datasets. Notably, the WERs for CMU_test and PF_br_test dropped to 1.86 and 3.10, respectively, nearing human-level accuracy. These results indicate that having a similar distribution of data improves performance on both seen and unseen child speech datasets. However, this improvement also points to a potential limitation: while the models became more proficient with data similar to their training set, their ability to generalize across diverse linguistic backgrounds weakened. The results underline the critical balance needed in selecting training datasets that perform well across a broad spectrum of speech recognition tasks.

6) GROUP F

In the Group F experiments, the finetuning included NN_10 along with the Group D training datasets. On comparing groups D and F, the addition of this small dataset of non-native speech resulted in significant improvements in the performance on all non-native child speech test datasets, while the performance on the other test datasets remained unchanged. The addition of NN_10 also led to a decrease in the WER on CMU_test. This demonstrates that whisper finetuning can enhance ASR performance on non-native child speech in a low-resource scenario and can be extended to other multi-accented non-native child speech.

7) GROUP G

The Group G finetuning experiments substituted NN_10 with NN_20, compared to Group F. This adjustment led to additional improvements across all non-native test datasets, with WERs dropping by 1-3% for each test dataset relative to the results from Group F. This further shows the benefit of including more extensive non-native speech data in finetuning to enhance ASR performance on non-native child speech.

8) GROUP H

In the Group H experiments, the finetuning included NN_10 along with the Group E training datasets. CMU_train was included in the finetuning to see its impact when used in conjunction with the non-native datasets. By looking at the average WERs, no significant difference between groups F and H can be observed, except for the CMU_test WER, which was expected. Surprisingly, adding CMU_train in finetuning didn't impact the performance on non-native test datasets in this group.

9) GROUP I

In the Group I finetuning experiments, the NN_20 training data was included instead of NN_10 in Group H. This resulted in further improvements, as WERs decreased by between 1-3% on all test datasets compared to Group H. Furthermore, it can be noted that the inclusion of CMU_train in the finetuning process did not have any noticeable effect on performance on non-native test datasets within this group.

B. DISCUSSION

The findings of this study offer significant insights into the feasibility of finetuning the whisper model on different combinations of child audio data to improve child speech recognition in both native and non-native English accents. This section delves deeper into the nuances of model performance, emphasizing the influence of model size, the role of dataset-specific finetuning, and the model's capability to adapt to diverse linguistic environments. The analysis in this section also addresses the challenges in accent recognition and evaluates the whisper model's resilience to catastrophic forgetting, highlighting its potential in the evolving field of speech recognition technology.

1. Model Size: Smaller models (e.g., Tiny, Small) tend to have higher WER scores compared to larger models (e.g., Medium, Large-V2). This suggests that larger models have a better capacity to capture and represent speech patterns, which leads to lower WER scores at the inference stage. It can also be seen that there is only a 1-2% WER difference between medium, medium.en, and large-V2 models, suggesting an upper limit to the generalizability of models with an increase in model size.

2. No Finetuning: For the Group-A experiments, it can be observed that on American English test datasets, such as MyST_test and CMU_test, the models generally had low WERs without any finetuning, as compared to model results on other test datasets. This implies that American-accented child speech has acoustic properties similar to adult speech.

3. Generalization: The models trained on the MyST_train dataset (Groups B through I) generally exhibit good generalization to other test scenarios, exhibiting relatively low WER. This suggests that the finetuned models can effectively adapt to diverse speech recognition tasks even when trained with a single-child speech dataset.

4. Finetuning Impact: Finetuning the models on specific datasets (Groups B through I) consistently leads to improved performance compared to the models without finetuning (Group A). This highlights the importance of adapting models to domain-specific data for better speech recognition.

5. Dataset Contribution: Among the finetuning datasets, the PF_br_train dataset (used in groups D, F, and G)

consistently provides the most significant improvements in WER scores across various test scenarios. It indicates that incorporating a dataset with diverse linguistic features can greatly benefit the model's performance.

6. Limited impact of the CMU_train dataset: Finetuning with the CMU_train dataset (Groups C and E) shows relatively smaller improvements on test datasets as compared to PF_br_train. This suggests that the CMU_train dataset might not capture linguistic features as effectively as the PF_br_train dataset.

7. Additional Data Impact: The inclusion of additional multi-accented non-native training datasets represented by NN_10 and NN_20 (used in Groups F, G, H, and I) yields substantial improvements in WER scores on non-native child speech test datasets (Group B, C, D, and E). This implies that additional non-native training datasets, even with amounts as low as 10% of the unseen dataset can improve the ASR system's performance.

8. Language and Accent: In our experiments, various child speech accents were used. Among the accented speech, British, Italian, and American accents are easier to improve on for child ASR tasks. German and Chinese accents still posed challenges in ASR accuracy, although small improvements were still seen.

9. Catastrophic Forgetting: The WER on adult speech (LibriTTS 'dev-clean') remained in the range of 4-6% for all finetuning experiments. This shows that whisper doesn't suffer from the catastrophic forgetting problem [36], which appears when a model is retrained with a different dataset than the original dataset and the effect being a significant reduction in performance data from the original training domain. Whisper models were able to retain a similar WER accuracy for adult speech while also improving the WER for child speech ASR. This may be attributed to careful training strategies, architectural design, and regularization techniques used in whisper [4].

C. COMPARISON WITH PREVIOUS SOTA RESULTS

Table 4 compares the results we obtained on the various test sets with previously reported results in the literature. Our results show significant improvements over the previously reported results. It is important to note, however, that prior researchers employed varied methodologies for data cleaning, and in the absence of a uniform standard for this process, a direct comparison cannot be provided. Consequently, we include these comparisons primarily to illustrate the effectiveness of our methodology on its own merits, rather than as a direct benchmark against previous work. This approach allows us to highlight the significant improvements our research contributes to the field while acknowledging the methodological differences that exist in data preprocessing practices. We report relative WER improvements between 29.7% on the MyST_test, 41.5% on the PF_br_test, 89.1% on the CMU_test, and 85.1% on the PF_sw_test datasets. During our research, other similar studies with whisper finetuning were also conducted which utilized different volumes of the MyST dataset for finetuning and testing. The WER from these studies are also presented in Table 4 (marked in blue) to draw a comparison of whisper finetuning with varying volumes of the same child speech dataset.

TABLE 4
COMPARISON OF PREVIOUSLY REPORTED WER RESULTS WITH OUR RESULTS

Test Data	Approach [training data]	WER	Relative WER improve-ment
MyST _test	**-Ours [MyST_train:55 hrs]** -DRAFT: Self-supervised [240hrs] [37] -Whisper-Medium [55hrs] [42] -Whisper-Medium [125hrs] [42]	**11.62** 16.53 14.40 8.61	29.7% over non-whisper models
PF_br _test	**-Ours [PF_br_train: 10 hrs]** -Filter-based discriminative autoencoder [8.4 hrs] [38] -wav2vec2-SSL [7.4 hrs] [17]	**2.84** 18.77 4.86	41.5% over previously reported lowest WER
CMU _test	**-Ours [CMU_train: 7 hrs]** -TDNN-F [54.90 hrs] [39], -HMM-DNN [6.34 hrs] [18] -TDNN-HMM [6.34 hrs] [40] -Encoder-Decoder VC [7.28 hrs] [41]	**1.75** 16.00 19.67 19.80 21.51	89.1% over previously reported lowest WER
PF_sw _test	**-Ours [PF_sw_train: 0.24 hrs]** -HMM-DNN [4 hrs] [18]	**7.66** 51.58	85.1%
PF_ge _test	**-Ours [PF_ge_train: 0.68 hrs]**	**34.26**	NA
PF_it_ test	**-Ours [PF_it_train: 0.7 hrs]**	**4.11**	NA
SO_te st	**-Ours [SO_train: 0.48 hrs]**	**14.31**	NA

Note: These results are provided to show comparisons on the same datasets. Dataset distributions used for training/testing will vary in these papers. We used an 80:20 split for testing and training. This approach uses only a small percentage of data for training as compared to other papers mentioned. We did not find previously reported results on PF_ge_test, PF_it_test, and SO_test datasets.

VI. CONCLUSION

This study aims to enhance the performance of ASR on native and non-native English child speech datasets available for research use. Whisper models, pretrained with huge amounts of data, are selected as the basis of the experimental studies conducted as part of this work. Our work adapts these pretrained whisper models to non-native child speech using finetuning. The effectiveness of whisper finetuning on ASR performance is studied through various experimental combinations of datasets. It was observed that whisper finetuning improves the ASR performance on non-native English children's speech, a low-resource domain.

Additionally, our approach outperforms previously reported results on the non-native child speech datasets used in this paper by using only 10% distributions of these datasets during finetuning. Best WERs on Swedish, German, Italian, and Chinese accented non-native English child speech are reported in this paper. Using child speech data with different linguistic features can benefit the overall ASR performance. German-accented speech was the most challenging for ASR while British and American English speech was the least challenging. It was also observed that Whisper does not suffer from catastrophic forgetting when finetuning on new datasets.

For future work, we aim to include more training datasets from other low-resource languages in finetuning to further improve on these baseline results. The influence of external language models at the decoding stage on non-native child speech will also be examined. We also intend to conduct a three-way experimental analysis with wav2vec2 [3], Whisper [4], and Conformer [2] models to study the strengths and limitations associated with each model for working with child speech.

ACKNOWLEDGMENT

The authors would like to acknowledge experts from Xperi Ireland: Gabriel Costache, Zoran Fejzo, and George Sterpu for providing their expertise and feedback.

REFERENCES

[1] Kriman, Samuel, et al. "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[2] Gulati, Anmol, et al. "Conformer: Convolution-augmented transformer for speech recognition." *arXiv preprint arXiv:2005.08100* (2020).

[3] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." *arXiv preprint arXiv:2212.04356* (2022).

[5] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999, doi: 10.1121/1.426686.

[6] I. Chandra Yadav and G. Pradhan, "Pitch and noise normalized acoustic feature for children's ASR," *Digital Signal Processing: A Review Journal*, vol. 109, p. 102922, 2021, doi: 10.1016/j.dsp.2020.102922.

[7] S. Lee, A. Potamianos, and S. S. Narayanan, "Analysis of children's speech: duration, pitch and formants," in *EUROSPEECH*, 1997.

[8] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition," pp. 135–140, 2014, doi: 10.1109/SLT.2014.7078563ï.

[9] Claus, Felix, Hamurabi Gamboa Rosales, Rico Petrick, Horst-Udo Hain, and Rüdiger Hoffmann. "A survey about databases of children's speech." In *INTERSPEECH*, pp. 2410-2414. 2013.

[10] P. Vieting, C. Lüscher, J. Dierkes, R. Schlüter, and H. Ney, "Efficient Use of Large Pre-Trained Models for Low Resource ASR," Oct. 2022, doi: 10.48550/arxiv.2210.15445.

[11] J. Bai *et al.*, "Joint unsupervised and supervised training for multilingual ASR," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, pp. 6402–6406, 2022, doi: 10.1109/ICASSP43922.2022.9746038.

[12] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing ASR Pretraining for Low-Resource Speech-to-Text Translation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May, pp. 7909–7913, May 2020, doi: 10.1109/ICASSP40776.2020.9053847.

[13] S. Squartini, M. Scarpiniti, J.-T. Chien, J. Camilo Vásquez-Correa, and A. Á. Muniain, "Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper," *Sensors 2023, Vol. 23, Page 1843*, vol. 23, no. 4, p. 1843, Feb. 2023, doi: 10.3390/S23041843.

[14] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards Better Domain Adaptation for Self-Supervised Models: A Case Study of Child ASR," *IEEE Journal on Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, Oct. 2022, doi: 10.1109/JSTSP.2022.3200910.

[15] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-Native Children Speech Recognition Through Transfer Learning," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 6229–6233, Sep. 2018, doi: 10.1109/ICASSP.2018.8462059.

[16] P. Gurunath Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer Speech & Language*, vol. 63, p. 101077, Sep. 2020, doi: 10.1016/J.CSL.2020.101077.

[17] Thienpondt, Jenthe, and Kris Demuynck. "Transfer Learning for Robust Low-Resource Children's Speech ASR with Transformers and Source-Filter Warping." *Proc. Interspeech 2022*.

[18] Rolland, Thomas, Alberto Abad, Catia Cucchiarini, and Helmer Strik. "Multilingual Transfer Learning for Children Automatic Speech Recognition." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7314-7320. 2022.

[19] T.-H. Lo, F.-A. Chao, S.-Y. Weng, and B. Chen, "The NTNU System at the Interspeech 2020 Non-Native Childrens Speech ASR Challenge."

[20] M. Shahin, R. Lu, J. Epps, and B. Ahmed, "UNSW System Description for the Shared Task on Automatic Speech Recognition for Non-Native Children's Speech," 2020, doi: 10.21437/Interspeech.2020-3111.

[21] G. Xu, S. Yang, L. Ma, C. Li, and Z. Wu, "The TAL system for the INTERSPEECH2021 Shared Task on Automatic Speech Recognition for Non-Native Childrens Speech," 2021, doi: 10.21437/Interspeech.2021-1104.

[22] P. Gurunath Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, Mar. 2022, doi: 10.1016/j.csl.2021.101289.

[23] Gerosa, Matteo, et al. "A review of ASR technologies for children's speech." *Proceedings of the 2nd Workshop on Child, Computer and Interaction*. 2009.

[24] A. Narayanan *et al.*, "Toward Domain-Invariant Speech Recognition via Large Scale Training," *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*, pp. 441–447, Feb. 2019, doi: 10.1109/SLT.2018.8639610.

[25] W. Chan, D. S. Park, C. A. Lee, Y. Zhang, Q. v Le, and M. Norouzi, "SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network."

[26] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

[27] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *7th International Conference on Learning Representations, ICLR 2019*, Nov. 2017, doi: 10.48550/arxiv.1711.05101.

[28] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." In *International conference on machine learning*, pp. 1310-1318. Pmlr, 2013.

[29] Ward, Wayne, Ron Cole, and Sameer Pradhan. "My science tutor and the myst corpus." *Boulder Learning Inc* (2019).

[30] A. Batliner *et al.*, "The PF STAR Children's Speech Corpus", doi: 10.21437/Interspeech.2005-705.

[31] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids speech corpus," *Corpus of children's read speech digitized and transcribed on two CD-ROMs, with assistance from Multicom Research and David Graff. Published by the Linguistic Data Consortium, University of Pennsylvania*, 1997.

[32] J. Zhang *et al.*, "speechocean762: An Open-Source Non-native English Speech Corpus For Pronunciation Assessment," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 6, pp. 4386–4390, Apr. 2021, doi: 10.48550/arxiv.2104.01378.

[33] H. Zen *et al.*, "LibriTTS: A corpus derived from libri speech for text-to-speech," *arXiv*. 2019. doi: 10.21437/interspeech.2019-2441.

[34] R. Jain, A. Barcovschi, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A Wav2vec2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition," Apr. 2022, doi: 10.48550/arxiv.2204.05419.

[35] Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., Cucu, H. (2023) Adaptation of Whisper models to child speech recognition. Proc. INTERSPEECH 2023, 5242-5246, doi: 10.21437/Interspeech.2023-935.

[36] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A. and Hassabis, D., 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), pp.3521-3526.

[37] R. Fan and A. Alwan, "DRAFT: A Novel Framework to Reduce Domain Shifting in Self-supervised Learning and Its Application to Children's ASR," 2022, doi: 10.21437/Interspeech.2022-11128.

[38] C.-L. Tai, H.-S. Lee, Y. Tsao, and H.-M. Wang, "Filter-based Discriminative Autoencoders for Children Speech Recognition." *arXiv preprint arXiv:2204.00164* (2022).

[39] F. Wu, L. Paola Garcia, D. Povey, and S. Khudanpur, "Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network," 2019, doi: 10.21437/Interspeech.2019-2980.

[40] Z. Fan, X. Cao, G. Salvi, and T. Svendsen, 'Using Modified Adult Speech as Data Augmentation for Child Speech Recognition', in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[41] P. N. Sudro, A. Ragni, and T. Hain, 'Adapting Pretrained Models for Adult to Child Voice Conversion', in 2023 31st European Signal Processing Conference (EUSIPCO), 2023, pp. 271–275.

[42] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, 'Kid-Whisper: Towards Bridging the Performance Gap in Automatic Speech Recognition for Children VS. Adults', arXiv [eess.AS]. 2023.

RISHABH JAIN received the B.Tech. degree in computer science and engineering from Vellore Institute of Technology (VIT), in 2019, and the M.S. degree in data analytics from the University of Galway, Ireland, in 2020. He is currently working as a Research Assistant at the University of Galway under the DAVID (Data-center Audio/Visual Intelligence on-Device) project. He is also pursuing a PhD degree from the University of Galway, Ireland. His research interests include machine learning and artificial intelligence specifically in the domain of speech understanding, text-to-speech, speaker recognition and automatic speech recognition.

ANDREI BARCOVSCHI received his B.Eng. degree in Electronic and Computer Engineering from the University of Galway (prior to 2023: National University of Ireland Galway, NUIG) in 2020 and his M.Sc. in Artificial Intelligence from NUIG in 2021. He is currently in his first year of a Ph.D. degree in Artificial Intelligence at the University of Galway, researching speech synthesis and conversion technologies, text-to-speech, and speech-to-text. He is interested in a broad range of machine learning and artificial intelligence topics.

MARIAM YAHAYAH YIWERE received her Bachelor of Science degree from the Department of Computer Science at the Kwame Nkrumah University of Science and Technology in Kumasi, Ghana in 2012. She received her Master of Engineering degree and a PhD degree from the Department of Computer Engineering, Hanbat National University, South Korea in August 2015, and February 2020 respectively. Since October 2020, Mariam has been working on the DTIF/DAVID project as a postdoctoral researcher at the College of Science and Engineering, University of Galway, Ireland. Her research interests include Text-to-Speech Synthesis, Speaker Recognition and Verification, Sound Source Localization, Deep Learning, and Computer Vision.

PETER CORCORAN (Fellow, IEEE) currently holds the Personal Chair in electronic engineering with the College of Science and Engineering, University of Galway, Ireland. He was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He has more than 600 cited technical publications and patents, more than 120 peer-reviewed journal articles, and 160 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is also a member of the IEEE Consumer Technology Society for more than 25 years and the Founding Editor of IEEE Consumer Electronics Magazine.

HORIA CUCU received the B.S. and M.S. degrees in applied electronics from the University Politehnica of Bucharest (UPB), Romania, in 2008 and the Ph.D. degree in electronics and telecom from the same university in 2011.

From 2010 to 2017, he was a Teaching Assistant and then Lecturer at UPB. He currently serves as an Associate Professor at the same university. In this position, he authored over 75 scientific papers in international conferences and journals, served as project director for 7 research projects, and contributed as a researcher to 10 other research grants. He holds two patents. In addition, he founded and led Zevo Technology, a speech start-up dedicated to integrating state-of-the-art speech technologies in various commercial applications. His research interests include machine/ deep learning and artificial intelligence, with a special focus on automatic speech and speaker recognition, text-to-speech synthesis, and speech emotion recognition. Dr. Cucu was awarded the Romanian Academy prize “Mihail Drăgănescu” (2016) for outstanding research contributions in Spoken Language Technology, after developing the first large-vocabulary automatic speech recognition system for the Romanian language.

Appendix F

A comparative analysis between Conformer Transducer, Whisper and Wav2vec2 for improving the child speech recognition.

Authors: Andrei Barcovschi (AB), Rishabh Jain (RJ), and Peter Corcoran (PC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	AB: 65%, RJ: 25%, PC:10%
Experiments and Implementation	AB: 70%, RJ: 30%
Background	AB: 60%, RJ: 40%
Manuscript Preparation	AB: 60%, RJ: 30%, PC:10%

A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition

Andrei Barcovschi
Ph.D. Student
University of Galway
Galway, Ireland
a.barcovschi1@universityofgalway.ie

Rishabh Jain
C3 Imaging Research Center
University of Galway
Galway, Ireland
rishabh.jain@universityofgalway.ie

Peter Corcoran
C3 Imaging Research Center
University of Galway
Galway, Ireland
peter.corcoran@universityofgalway.ie

2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) | 979-8-3503-2797-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/SpeD59241.2023.10314867

***Abstract*— Automatic Speech Recognition (ASR) systems have progressed significantly in their performance on adult speech data; however, transcribing child speech remains challenging due to the acoustic differences in the characteristics of child and adult voices. This work aims to explore the potential of adapting state-of-the-art Conformer-transducer models to child speech to improve child speech recognition performance. Furthermore, the results are compared with those of self-supervised wav2vec2 models and semi-supervised multi-domain Whisper models that were previously finetuned on the same data. We demonstrate that finetuning Conformer-transducer models on child speech yields significant improvements in ASR performance on child speech, compared to the non-finetuned models. We also show Whisper and wav2vec2 adaptation on different child speech datasets. Our detailed comparative analysis shows that wav2vec2 provides the most consistent performance improvements among the three methods studied.**

Keywords*— *Child Speech Recognition, Automatic Speech Recognition, Conformer-transducer, wav2vec2, Whisper model, MyST, PF-STAR, CMU_Kids

I. Introduction

In the domain of Automatic Speech Recognition (ASR), several challenges persist, such as limited training data, untranscribed data, and difficulty in low-resource languages and children's speech. Recent research efforts have addressed some of these issues, leading to impressive ASR performance for adult speech, even achieving human-level performance [1]–[5]. However, progress in ASR for child speech has been slower, primarily due to the scarcity of annotated child-speech datasets required for effective training. Child speech datasets are challenging to collect and annotate, unlike adult speech data (as discussed in [6]). Moreover, inherent differences between adult and child voices, including pitch, linguistic and acoustic features, and pronunciation ability [7], [8], further hinder the performance of ASR models on child speech. The shorter vocal tract length and higher fundamental frequency [9] of children's voices also contribute to the complexity of accurately recognizing child speech.

The advantages and disadvantages of supervised and unsupervised ASR training approaches have been observed in recent developments, particularly in the context of child speech recognition. Unsupervised pretraining techniques like wav2vec2 [3] have shown significant improvements in child ASR [10]–[12]. However, their reliance on a finetuning stage with labeled data can limit their usefulness as they may overfit to specific datasets and not generalize well to diverse distributions. On the other hand, supervised learning approaches in child ASR [13]–[15] have explored transfer learning from adult to child speech [10], [13], [16], data augmentation methods [17]–[19], and weakly supervised training [15], [16], [20]. Recent findings [21], [22] indicate that supervised methods, involving pretraining on multiple datasets/domains, can enhance model robustness and generalization performance on unseen datasets. Nevertheless, each approach presents its trade-offs in terms of adaptability and scalability for diverse real-world speech recognition scenarios.

In this work, we use recent State-of-the-Art (SOTA) ASR models, Conformer-transducer for the task of child speech recognition. We also provide a comparative analysis of this model with our previously benchmarked results on wav2vec2 [23] and whisper [24]. Whisper is a supervised learning-based ASR system, which uses large amounts of labeled audio data. It uses weakly supervised pretraining beyond English-only speech recognition to be multilingual and multitask, showing great performance on different multilingual adult speech datasets [4]. The wav2vec2 is a self-supervised pretraining method for speech representations, leading to data-efficient finetuning for downstream ASR tasks. Conformer-transducer, combining CNNs and Transformers for end-to-end speech recognition, offers streaming capabilities and efficient long-range dependency modeling. While wav2vec2 is data-efficient and Whisper and Conformer-transducer excel in real-time processing, each model has unique strengths, making the choice dependent on factors like performance, model size, and application requirements. Since these models perform well on adult speech and gave SOTA results on widely used adult speech datasets, it was decided to use these models on different child speech datasets. We also finetune these models using different combinations of child speech datasets to see the subsequent speech recognition performance on different seen and unseen distributions of child speech datasets [25]–[27]. Our goal is to evaluate the efficacy of these methodologies in child speech analysis and determine their potential for enhancing child ASR technology and developing educational tools for children.

II. Model Description

A. Conformer-transducer [2]

The Conformer-transducer ASR model combines the benefits of both the transformer and CNN into a single architecture, namely the efficient global-level modeling of long-range dependencies in audio samples introduced by self-attention, and the finer-grained modeling of local dependencies enabled by convolutional kernels, respectively. The encoder network consists of a stack of Conformer blocks replacing the Transformer blocks [28]. A Conformer block consists of a feed-forward module followed by a multi-headed self-attention module, a convolution module, and finally

another feed-forward module. Half-step residual connections always follow the feed-forward modules and a Layernorm is added as the last step in each block. The architecture of the Conformer encoder can be seen in Figure 1.

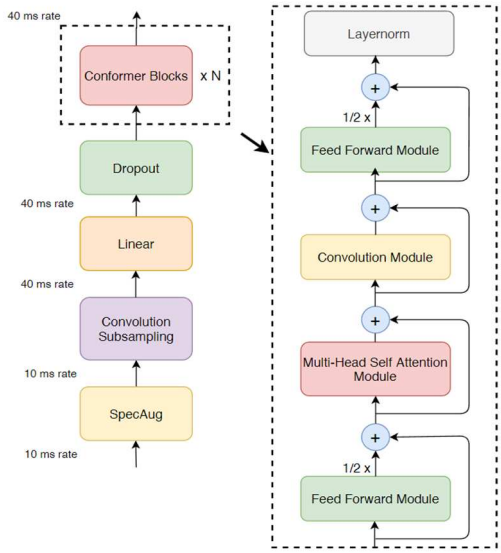


Figure 1: Conformer encoder model architecture [2].

Conformer-transducer models offer an improvement in WER for adult speech over the RNN-T and the Transformer-transducer architectures [2]. The Conformer-transducer uses the autoregressive transducer decoder, dropping the original simpler LSTM decoder. For the task of ASR, using the transducer decoder and transducer loss instead of the Connectionist Temporal Classification (CTC) [29] reduces incorrect spellings due to its autoregressive property, which implicitly models the inherent dependency between predicted output tokens, while CTC assumes that the output tokens are conditionally independent. However, this comes at the cost of larger GPU memory requirements for training and slower decoding speeds. Using a transducer approach introduces two new networks – the Decoder and the Joint model. The output of the Conformer's Encoder is inputted to the joint model, along with the autoregressive decoder model's output, and autoregressively produces a joint probability distribution over the known token vocabulary. At training time, the transducer loss is calculated over the output of the joint network.

B. *Whisper [4]*

Whisper represents a significant advancement in weakly supervised pre-training, extending its capabilities to encompass multilingual and multitask scenarios beyond English-only speech recognition. Its strength lies in a vast dataset comprising 680,000 hours of labeled audio, where 117,000 hours cover 96 different languages, and an additional 125,000 hours include X→en translation data, where X is a non-English language and 'en' represents English translated data. Employing a transformer-based architecture with residual connections, the model handles an entire speech processing pipeline, encompassing transcription, translation, voice activity detection, alignment, and language identification. The Whisper model operates on 80-channel log-Mel spectrograms, with the encoder-decoder Transformer network featuring two convolutional layers, sinusoidal positional encoding, and a stacked set of Transformer blocks. The decoder uses learned positional embeddings and the same number of Transformer blocks as in the encoder. A comprehensive explanation of the Whisper architecture is available in [4].

C. *wav2vec2 [3]*

wav2vec 2.0 is a speech recognition model based on self-supervised learning of speech representations through a two-stage architecture for pretraining and finetuning. The architecture comprises three key components: a CNN feature extractor, a transformer-based encoder, and a quantization module (see [3] for detailed information). During pretraining, the model is trained on a vast dataset of unlabeled speech data to acquire meaningful representations by capturing the temporal and spectral characteristics of speech. This is accomplished using a masked contrastive loss function. In the finetuning phase, the pretrained model is further trained on a smaller labeled dataset tailored for a specific downstream task. Here, the last layer of the pretrained model is substituted with a task-specific feed-forward layer, and the entire model is finetuned by minimizing the CTC loss [29] for ASR.

D. *Training Details*

All models were trained on A6000 GPUs with 48GB of memory. The architectural parameters for Whisper, wav2vec2, and Conformer-transducer models utilized in this study are detailed in Table I.

TABLE I. ARCHITECTURE PARAMETERS FOR CONFORMER-TRANSDUCER[2], WHISPER[4], AND WAV2VEC2 [3] MODELS

Models	Layers	Width	Heads	Learning Rate	Para-meters
Conformer-Transducer Models:					
Small	16	176	4	3.0	14M
Medium	16	256	4	3.0	32M
Large	17	512	8	3.0	120M
XLarge	24	1024	8	3.0	600M
Whisper Models:					
Tiny	4	384	6	1.5×10^{-3}	39M
Base	6	512	8	1×10^{-3}	72M
Small	12	768	12	5×10^{-4}	244M
Medium	24	1024	16	2.5×10^{-4}	769M
Large	32	1280	20	1.75×10^{-4}	1550M
wav2vec2 Models:					
Base	12	768	8	5×10^{-4}	95M
Large	24	1024	16	3×10^{-4}	317M

III. CORPUS DESCRIPTION

The Conformer-transducer pretrained models are trained on several thousand hours of English speech from diverse resources such as Librispeech, Fisher Corpus, Switchboard-1 Dataset, WSJ-0 and WSJ-1, National Speech Corpus, VCTK, VoxPopuli, Europarl, Multilingual Librispeech, Mozilla Common Voice, and People's Speech. The authors of Whisper [4] do not explicitly state the datasets used for training their models. Nonetheless, these trained models achieved SOTA results on various adult speech ASR datasets [4]. The wav2vec2-base model is pretrained with 960 hours of librispeech [30] and the wav2vec2-large model is pretrained with 60k hours of libri-light [31] datasets. In our study, we utilize three distinct child speech datasets and one adult speech dataset: MyST Corpus [25], PFSTAR dataset

[27], and CMU Kids dataset [26]. We maintain consistency with previous research wav2vec2 [23] and Whisper [24] to facilitate a direct comparison with the Conformer-transducer models.

A. Dataset cleanup

The cleaning process for the text labels involved removing abbreviations, punctuations, white spaces, and other non-alphanumeric characters, and converting all characters to lowercase. The audio data was modified to have a 16Khz sampling rate and 16-bit mono channel. For finetuning experiments, we used My Science Tutor (MyST) Corpus [25], an American English dataset. After cleaning and preparing this dataset according to [23], we divided 65 hours of clean child speech into two subsets: 55 hours for training and 10 hours for testing. Additionally, PFSTAR [27], a collection of words spoken by British English children, contributed 12 hours of audio, with 10 hours used for training and 2 hours used for testing. We also utilized CMU_Kids [26] corpus for validation-only, containing 9 hours of read-aloud sentences by children. While these datasets may not be extensive, they currently represent the best publicly available child speech datasets.

B. Dataset Usage

The datasets were divided according to their usage into a 'training' and an 'inference' set. This information is summarized in Table II.

TABLE II. DATASET USAGE

Usage	Dataset	Duration
Finetuning (Training)	MyST_55h	55 hours
	PFS_10h	10 hours
Inference (Testing)	dev-clean	9 hours
	MyST_test	10 hours
	PFS_test	2 hours
	CMU_test	9 hours

IV. EXPERIMENTS AND RESULTS

A. Codebase

The Whisper implementation used is provided here[1]. The fairseq[2] implementation of wav2vec2 is used for finetuning experiments. The relevant information regarding model training, hyperparameters, graphs/metrics, checkpoints, and dataset availability are made available on our GitHub[3]. As for Conformer, we use its Nvidia's implementation for our experiments[4].

B. Experiments

The first set of Conformer-transducer experiments involved evaluating the original publicly available models on different child audio evaluation datasets mentioned in Table II without finetuning. The model sizes used were Small, Medium, Large, and XLarge as mentioned in Table I. For Whisper experiments, we use the Tiny, Base, Small, Medium, Large, and Large-V2 models. There are two versions of each model: one trained with multilingual data and one specifically for the English language only (indicated by '.en' in the name). The detailed list of experiments is mentioned in [23]. For wav2vec2 experiments, we use the 'Base' and 'Large' models which are pretrained with 960 hours of Librispeech data [30] and 60,000 hours of Libri-light data [31] respectively. Two models with the best performance from the first set of experiments are selected for further finetuning, namely, the models with the lowest WER. Finetuning included three experimental configurations of training data: MyST_55h, PFSTAR_10h, and MyST_55h+PFSTAR_10h combined.

The Conformer-transducer finetuning experiments on child speech involved finetuning only the feed-forward layers of all the encoder's Conformer blocks along with all layers of the decoder and joint networks of the base models. This decision was taken based on selecting the best result from preliminary experiments that tested different training hyperparameters and the finetuning of different combinations of layers of the Conformer-transducer large model, which can be found in the Appendix. The Adam optimizer was used with a base learning rate of 3.0 in combination with the Noam learning rate scheduler which linearly increased the learning rate for the first 40,000 steps before decaying exponentially. Greedy batch decoding was used as the token decoding strategy and for all experiments a unigram-based sentence-piece tokenizer with a vocabulary size of 1024 tokens was created for each unique finetuning dataset combination. The models were finetuned up to 500 epochs.

For whisper and wav2vec2 finetuning, the finetuning setup was kept consistent with previously reported results on Whisper [24] and wav2vec2 [23] approaches to provide a fair comparative analysis. We use a learning rate of 1×10^{-5} for all Whisper finetuning experiments. The wav2vec2-base was finetuned with a learning rate of 1×10^{-4}, while wav2vec2-large was finetuned with a learning rate of 2.5×10^{-5}. Finetuning both approaches involves training the final layer of the models and freezing all others, as described by the respective authors. The Whisper model undergoes finetuning by minimizing the cross-entropy objective function, whereas wav2vec2 is finetuned by minimizing the CTC loss.

C. Results and Discussions

a) **No-Funetuning Experiments***:* Table III shows Word Error Rates (WERs) of original, non-finetuned Whisper, wav2vec2, and Conformer-transducer models on child speech evaluation datasets mentioned in Table II. No initial finetuning was performed over these models. A general trend of high WER on the MyST_test evaluation set can be observed across all the Whisper and Conformer-transducer models with most hovering around the 25% mark even for the much larger models. Only the wav2vec2 models perform better on MyST_test, displaying WERs that are approximately 10 points lower. We use these experiments as a baseline for further finetuning. The models with the lowest WER were chosen for providing executing further finetuning experiments with child speech.

[1] **Whisper Implementation:** https://github.com/huggingface/community-events/tree/main/whisper-fine-tuning-event

[2] **wav2vec2 Fairseq:** https://github.com/facebookresearch/fairseq/

[3] **GitHub:** https://github.com/C3Imaging/child_asr_conformer/

[4] **Conformer-transducer**: https://github.com/NVIDIA/NeMo/

TABLE III. *WER FOR DIFFERENT NON-FINETUNED WHISPER, WAV2VEC2, AND CONFORMER-TRANSDUCER MODELS ON CHILD SPEECH (MYST, PFSTAR, AND CMU-KIDS) EVALUATION DATASETS*

Name	Models	MyST_ test	PFS_ test	CMU_ test
Conformer-Transducer	Small	**21.34**	12.68	16.05
	Medium	24.99	11.58	17.51
	Large	25.91	8.94	15.06
	Xlarge	24.42	**8.22**	**14.83**
Whisper	Tiny	40.09	159.57	30.63
	Tiny.en	33.02	47.11	27.32
	Base	32.14	100.07	25.03
	Base.en	29.15	45.70	20.75
	Small	26.22	111.75	18.52
	Small.en	26.72	39.00	16.82
	Medium	25.11	80.97	**12.67**
	Medium.en	28.06	**35.25**	14.00
	Large	25.24	84.52	13.70
	Large-V2	**25.00**	73.68	12.69
wav2vec2	wav2vec2-base	15.41	11.20	16.33
	wav2vec2-large	**12.50**	**8.56**	**14.85**

The Conformer-transducer Small model, which is significantly smaller than Whisper's Tiny model and five times smaller than Whisper's Base model outperforms both significantly on all three child audio evaluation sets. The Conformer-transducer Medium model, comparable in size to Whisper's Tiny model also outperforms Whisper but does not reach the same accuracy as wav2vec2-base, which is three times bigger, though the WERs are relatively close. The large Conformer-transducer model again outperforms the Whisper model of roughly the same parameter size range (Whisper Small) across all three child evaluation datasets and performs better on PFS_test and CMU_test than wav2vec2-base while its performance on MyST_test is significantly worse than wav2vec2-base. Conformer-transducer Xlarge model, which is twice the size of wav2vec2-large, only performs on par with it for PFS_test and CMU_test, while again showing a much poorer result on MyST_test. XLarge model, being slightly smaller in size than the Whisper Medium model, slightly outperforms the Whisper Medium model on MyST_test, significantly outperforms the Whisper Medium model on PFS_test, and does not outperform on CMU_test.

Overall, it can be observed that smaller Conformer-transducer models perform better than their small Whisper and wav2vec2 counterparts, while with an increase in parameter size, the Whisper and wav2vec2 models tend to outperform Conformer-transducer equivalents, suggesting that the Conformer-transducer loses its generalization capabilities with an increase in parameter size. The Conformer-transducer 'Large' and 'Xlarge' models demonstrated competitive performance in most cases. The Whisper models generally exhibited higher WERs compared to the Conformer-transducer models. However, the 'Medium' and 'Large' Whisper models showed impressive results on all three datasets. The wav2vec2 models, particularly the 'wav2vec2-large' model, achieved the lowest WERs among all the models evaluated.

b) **Comparative analysis between Conformer-transducer, Whisper, and wav2vec2 after finetuning:** Conformer-transducer finetuning experiments involved using the Large and Xlarge models, selected after analyzing the results of non-finetuned models on child evaluation datasets. Whisper finetuning included the Medium.en and Large-V2 models while wav2vec2 finetuning involved wav2vec2-base and wav2vec2-large models. These models were finetuned on MyST_55h, PFSTAR_10h, and a combination of both datasets.

TABLE IV. *WER ON CHILD EVALUATION DATASETS FOR DIFFERENT WHISPER, WAV2VEC2, AND CONFORMER-TRANSDUCER MODELS FINETUNED ON MYST, PFSTAR, AND MYST+PFSTAR-COMBINED DATASETS*

Name	Models	MyST _test	PFS _test	CMU _test
MyST (55 Hours) Finetuning:				
Conformer-Transducer	Large	14.17	44.02	27.03
	XLarge	13.79	43.57	20.63
Whisper	Medium.en	**11.81**	17.83	**15.07**
	Large-V2	12.28	**10.88**	15.67
wav2vec2	wav2vec2-base	8.13	14.77	16.47
	wav2vec2-large	**7.51**	**12.46**	**15.25**
PFSTAR (10 Hours) Finetuning:				
Conformer-Transducer	Large	90.00	8.58	82.00
	XLarge	86.79	**6.31**	75.26
Whisper	Medium.en	15.84	3.14	15.53
	Large-V2	**15.79**	**2.88**	15.22
wav2vec2	wav2vec2-base	31.86	**3.48**	27.49
	wav2vec2-large	**27.17**	3.50	**21.35**
MyST (55 Hours) + PFSTAR (10 Hours) Finetuning:				
Conformer-Transducer	Large	13.86	4.44	25.00
	XLarge	13.61	4.3	21.21
Whisper	Medium.en	12.33	3.32	15.08
	Large-V2	13.34	4.17	17.11
wav2vec2	wav2vec2-base	7.94	**2.91**	15.97
	wav2vec2-large	**7.42**	2.99	**14.18**

A comparison between the Conformer-transducer, Whisper, and wav2vec2 WERs on the same evaluation sets can be seen in Table IV. First, a substantial increase in WER on the PFS_test and CMU_test is observed for the Conformer-transducer models finetuned on MyST_55h, while the WER on MyST_test is still higher than that for all Whisper and wav2vec2 models. Considering that CMU_test is the noisiest evaluation dataset, it is possible that, due to the higher WER of the Conformer-transducer on this set, the Conformer-transducer models deal worse with noisy datasets than the other model architectures. The results of the experiments with child speech finetuning show that wav2vec2 finetuning using MyST_55h resulted in lower WER compared to Whisper finetuning on MyST_test.

Finetuning the Conformer-transducer models on PFS_10h reduces the WER on PFS_test but again not to the same low levels as Whisper or wav2vec2 finetuning. Meanwhile, WERs on MyST_test and CMU_test is considerably higher for the Conformer-transducer models, again suggesting poor performance on noisier datasets. Finetuning the Conformer-transducer on a combination of MyST_55h and PFS_10h did not provide any improvements over the other models. However, when comparing to single dataset finetuning, the combined finetuning measurably improves the performance across all three evaluation datasets, suggesting that the model generalizes better when trained on more diverse and seen datasets.

Even though larger models tend to perform slightly better than their smaller counterparts, the performance gain from using larger models might not justify the additional computational cost and memory requirements, especially considering that the difference in WER between these models

is relatively small. The performance of the models is heavily influenced by the finetuning dataset. Models finetuned on the MyST dataset tend to perform better on the MyST_test evaluation dataset, while those fine-tuned on the PFSTAR dataset achieve better results on the PFS_test evaluation dataset. This suggests that domain-specific finetuning is crucial for achieving better performance on domain-specific evaluation datasets.

Overall, the results in Table IV indicate that wav2vec2 may be the best ASR model for finetuning on child data, as the models are smaller and require drastically less data to train than Whisper models which show slightly poorer or comparable results at best. It consistently outperforms the other models across different finetuning datasets and evaluation datasets and achieves the lowest WER values for both MyST and PFSTAR datasets and their combination, indicating its effectiveness in capturing relevant speech features and generalizing to unseen data. While wav2vec2 shows promising results, it is important to note that the table might not cover all possible scenarios and datasets. Further evaluation and testing on different datasets would be required to validate the model's robustness and generalization capabilities.

V. Conclusions

In this paper, the Conformer-transducer ASR model was compared against the Whisper and wav2vec2 models as approaches to improve the quality of child speech recognition. A fair comparison was conducted by ensuring that all models were evaluated within an identical parameter range and trained/evaluated using the same set of datasets. While the results show that finetuning the Conformer-transducer did not yield lower WER scores on child evaluation datasets compared to the Whisper or wav2vec2 finetuned models on the same datasets, there is still promise in using smaller-sized Conformer-transducer models for efficient low-resource deployment. The observed differences in finetuning performance may be attributed to the generalization capacity of the models, particularly for larger model sizes. It was evident that non-finetuned Conformer-transducer models had a more significant WER degradation compared to non-finetuned Whisper and wav2vec2 models as the model parameter size increased.

Furthermore, finetuned Conformer-transducer models perform worse on noisier evaluation datasets than Whisper and wav2vec2 models. Using a combination of datasets for finetuning improved WER scores across all datasets for the Conformer-transducer, suggesting that a more diverse finetuning dataset is needed for the model to generalize well to unseen data. On the other hand, when comparing non-finetuned models at smaller sizes, the Conformer-transducer model outperformed both the Whisper and wav2vec2 models within a similar parameter range across all child evaluation datasets. This indicates that Conformer-transducer models perform optimally at smaller sizes but may face challenges in maintaining generalization capabilities as their size increases. Overall, wav2vec2 showed the most promising results and can be considered to be the best ASR model for finetuning child data among the other models.

In future work, it is proposed to finetune the smaller Conformer-transducer models, namely Small and Medium, on child datasets. Additionally, more rigorous hyperparameter sweeping could provide lower WER scores as well as testing different decoding strategies such as beam-search with Time Synchronous Decoding (TSD) [32] or Alignment-Length Synchronous Decoding (ALSD) [33]. Finally, using different vocabulary sizes for the tokenizer may be investigated.

Acknowledgment

The authors would like to acknowledge experts from Xperi Ireland: Gabriel Costache, Zoran Fejzo, and George Sterpu for providing their expertise and feedback while working on this research.

Appendix

As of date, the only approach to finetuning Conformer-transducer models that are documented is simply training all layers of the encoder, decoder, and joint networks. However, we considered the possibility of finding a more optimal approach to finetuning which would lead to lower WER scores on the evaluation datasets. To determine the best combination of hyperparameters and what layers of the Conformer-transducer model to finetune for the main experiments detailed in Section IV.C, the large model was preliminarily fine-tuned on MyST_55h and the setup with the lowest WER on MyST_test was chosen as the finetuning approach to use.

The first approach involved finetuning all layers of all networks with the baseline hyperparameters recommended by the training scripts, which use the Adam optimizer with a learning rate of 5.0 and the Noam learning rate scheduler with 10,000 warmup steps. The lowest WER achieved on the MyST_test for this approach was 18.58%. The next approach modified the learning rate to 2.0, which led to a decreased WER of 16.3%. Further decreasing the learning rate to 1.0 achieved a WER of 14.55%. The next investigated approach involved finetuning just the feed-forward layers of the encoder network while freezing all other encoder layers, with a 1.0 learning rate and 10,000 Noam warmup steps, achieving a 14.21% WER. Using the Noam Hold learning rate scheduler with a warmup of 10,000 steps and a hold of 20,000 steps did not lead to improvements in WER on MyST_test. Finetuning only the final half of the feed-forward layers of the encoder instead of all the feed-forward layers also did not yield improvements. Finally, the best WER of 14.17% was achieved by finetuning all the feed-forward layers of the encoder with a learning rate of 3.0 and a Noam warmup of 40,000 steps. Note that all layers of the decoder and joint networks were fine-tuned in all of the preliminary experiments.

References

[1] S. Kriman *et al.*, 'Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions', in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6124–6128.

[2] A. Gulati *et al.*, 'Conformer: Convolution-augmented Transformer for Speech Recognition'.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, 'wav2vec 2.0: A Framework for Self-Supervised Learning of Speech

Representations', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 12449–12460.
[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, 'Robust speech recognition via large-scale weak supervision', in *International Conference on Machine Learning*, PMLR, 2023, pp. 28492–28518.
[5] D. Amodei *et al.*, 'Deep Speech 2: End-to-End Speech Recognition in English and Mandarin Baidu Research-Silicon Valley AI Lab *'.
[6] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, 'A survey about databases of children's speech.', in *INTERSPEECH*, 2013, pp. 2410–2414.
[7] S. Lee, A. Potamianos, and S. Narayanan, 'Acoustics of children's speech: Developmental changes of temporal and spectral parameters', *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999, doi: 10.1121/1.426686.
[8] S. Lee, A. Potamianos, and S. S. Narayanan, 'Analysis of children's speech: duration, pitch and formants', in *EUROSPEECH*, 1997.
[9] R. Serizel and D. Giuliani, 'Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition', in *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2014, pp. 135–140.
[10] J. Thienpondt and K. Demuynck, 'Transfer Learning for Robust Low-Resource Children's Speech ASR with Transformers and Source-Filter Warping'.
[11] R. Fan, Y. Zhu, J. Wang, and A. Alwan, 'Towards Better Domain Adaptation for Self-Supervised Models: A Case Study of Child ASR', *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1242–1252, 2022, doi: 10.1109/JSTSP.2022.3200910.
[12] R. Fan and A. Alwan, 'DRAFT: A Novel Framework to Reduce Domain Shifting in Self-supervised Learning and Its Application to Children's ASR'.
[13] T. Rolland, A. Abad, C. Cucchiarini, and H. Strik, 'Multilingual Transfer Learning for Children Automatic Speech Recognition', in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 7314–7320.
[14] F. Wu, L. Paola Garcia, D. Povey, and S. Khudanpur, 'Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network', 2019, doi: 10.21437/Interspeech.2019-2980.
[15] P. G. Shivakumar and S. Narayanan, 'End-to-end neural systems for automatic children speech recognition: An empirical study', *Comput. Speech Lang.*, vol. 72, p. 101289, 2022.
[16] Shivakumar, P.G. and Georgiou, P., 2020. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. Computer speech & language, 63, p.101077.
[17] K. Y. Chenpeng Du, 'Speaker Augmentation for Low Resource Speech Recognition', *ICASSP 2020 - 2020 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7719–7723, 2020.
[18] V. Kadyan, H. Kathania, P. Govil, and M. Kurimo, 'Synthesis Speech Based Data Augmentation for Low Resource Children ASR', *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, p. 12997, 2021, doi: 10.1007/978-3-030-87802-3_29.
[19] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, 'Data Augmentation Using CycleGAN for End-to-End Children ASR', *Eur. Signal Process. Conf.*, vol. 2021-August, pp. 511–515, 2021, doi: 10.23919/EUSIPCO54536.2021.9616228.
[20] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, 'A review of ASR technologies for children's speech', in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, 2009, pp. 1–8.
[21] A. Narayanan *et al.*, 'Toward Domain-Invariant Speech Recognition via Large Scale Training', *2018 IEEE Spoken Language Technology Workshop SLT 2018 - Proc.*, pp. 441–447, Feb. 2019, doi: 10.1109/SLT.2018.8639610.
[22] W. Chan, D. S. Park, C. A. Lee, Y. Zhang, Q. V. Le, and M. Norouzi, 'SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network'.
[23] R. Jain, A. Barcovschi, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, 'A Wav2vec2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition',Apr. 2022, doi: 10.48550/arxiv.2204.05419.
[24] Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., Cucu, H. (2023) Adaptation of Whisper models to child speech recognition. Proc. INTERSPEECH 2023, 5242-5246, doi: 10.21437/Interspeech.2023-935
[25] Ward, W., Cole, R., & Pradhan, S. (2019). My science tutor and the myst corpus. Boulder Learn. Inc.
[26] M. Eskenazi, J. Mostow, and D. Graff, 'The CMU kids speech corpus', *Corpus Child. Read Speech Digit. Transcribed Two CD-ROMs Assist. Multicom Res. David Graff Publ. Linguist. Data Consort. Univ. Pa.*, 1997.
[27] Russell, Martin. "The pf-star british english childrens speech corpus." The Speech Ark Limited (2006).
[28] A. Vaswani *et al.*, 'Attention is all you need', *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[29] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, 'Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks', in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, 'Librispeech: An ASR corpus based on public domain audio books', in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.
[31] J. Kahn *et al.*, 'Libri-light: A benchmark for asr with limited or no supervision', in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7669–7673.
[32] X. L. Aubert, 'An overview of decoding techniques for large vocabulary continuous speech recognition', *Computer Speech Language*, vol. 16, no. 1, pp. 89–114, 2002.
[33] G. Saon, Z. Tüske, and K. Audhkhasi, 'Alignment-length synchronous decoding for RNN transducer', in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7804–7808.

Appendix G

Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale.

Authors: Mariam Yiwere (MY), Andrei Barcovschi (AB), Rishabh Jain (RJ), Horia Cucu (HC) and Peter Corcoran (PC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	MY: 65%, AB: 10%, RJ: 10%, HC:10%, PC: 5%
Experiments and Implementation	MY: 60%, AB: 20%, RJ: 20%
Background	MY: 70%, AB: 15%, RJ: 15%
Manuscript Preparation	MY: 60%, AB: 10%, RJ: 10%, PC: 10%, HC: 10%

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 July 2023, accepted 22 August 2023, date of publication 20 September 2023, date of current version 10 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317360

RESEARCH ARTICLE

Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale

MARIAM YAHAYAH YIWERE[1], ANDREI BARCOVSCHI[1], RISHABH JAIN[1], (Graduate Student Member, IEEE), HORIA CUCU[2], (Member, IEEE), AND PETER CORCORAN[1], (Fellow, IEEE)

[1]School of Electrical and Electronics Engineering, University of Galway, Galway, H91 TK33 Ireland
[2]Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, 060042 Bucharest, Romania

Corresponding author: Mariam Yahayah Yiwere (mariam.yiwere@universityofgalway.ie)

This work was supported in part by the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project (2020–2023) funded by the Disruptive Technologies Innovation Fund (DTIF), in part by the College of Science and Engineering Ph.D. Research Scholarship with the University of Galway, and in part by the Science Foundation Ireland (SFI) ADAPT Center for Digital Media Research under Grant 13/RC/2106_P2.

ABSTRACT Technologies such as Text-To-Speech (TTS) synthesis and Automatic Speech Recognition (ASR) have become important in providing speech-based Artificial Intelligence (AI) solutions in today's AI-centric technology sector. Most current research work and solutions focus largely on adult speech compared to child speech. The main reason for this disparity can be linked to the limited availability of children's speech datasets that can be used in training modern speech AI systems. In this paper, we propose and validate a speech augmentation pipeline to transform existing adult speech datasets into synthetic child-like speech. We use a publicly available phase vocoder-based toolbox for manipulating sound files to tune the pitch and duration of the adult speech utterances making them sound child-like. Both objective and subjective evaluations are performed on the resulting synthetic child utterances. For the objective evaluation, the similarities of the selected top adults' speaker embeddings are compared before and after the augmentation to a mean child speaker embedding. The average adult voice is shown to have a cosine similarity of approximately 0.87 (87%) relative to the mean child voice after augmentation, compared to a similarity of approximately 0.74 (74%) before augmentations. Mean Opinion Score (MOS) tests were also conducted for the subjective evaluation, with average MOS scores of 3.7 for how convincing the samples are as child-speech and 4.6 for how intelligible the speech is. Finally, ASR models fine-tuned with the augmented speech are tested against a baseline set of ASR experiments showing some modest improvements over the baseline model finetuned with only adult speech.

INDEX TERMS Adult speech datasets, child speech datasets, synthetic child speech, speech data augmentation, CLEESE, speaker embeddings, pitch tuning, fundamental frequency.

I. INTRODUCTION

In recent years, rapid advances in Machine Learning (ML) and Deep Neural Network (DNN) techniques, together with tremendous increases in computational power, have led to a significant boost in the development of speech related technologies such as ASR [1], [2], [3], [4], [5], [6], [7], [8], TTS [9], [10], [11], [12], [13], [14] and Speaker Recognition [15], [16], [17], [18], [19] for multiple application domains. However, most of the solutions to date focus largely on adult speech, leading to poor performance when dealing with children's speech. The relatively smaller amount of work done specifically with child speech [20], [21], [22], [23], [24], [25], [26] has encountered significant challenges,

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik.

and as a result, children cannot fully benefit from some modern speech technologies. One of the main challenges is the limited availability of children's speech datasets [27], [28], [29], [30] necessary and suitable for training speech AI models. This is in contrast to adult speech where there is an abundance of large, publicly available and well-annotated datasets [31], [32], [33], [34], [35], [36]. These are harvested from the vast amounts of high-quality public data available online, including YouTube videos and professionally recorded audiobooks. Unfortunately, these adult datasets are not suitable for developing child-friendly speech AI solutions due to the innate differences between child and adult speech.

Child speech differs in multiple ways from adult speech owing essentially to the anatomical and morphological differences in their vocal-tract structure. Children have shorter vocal cords, giving their voices higher fundamental and formant frequencies compared to adults. In addition, children may have less control over articulation and non-linguistic aspects of speech such as prosody and therefore child speech exhibits higher spectral and temporal variation than adult speech [39].

On average, children also have slower speaking rates due to having longer phoneme durations [40]. They also exhibit higher pitch values: typically above 250Hz compared to average pitch values of 130Hz for adult males and 220Hz for adult females [41], [42]. For these reasons, it is important to gather and prepare good quality children's speech data to successfully train child-friendly speech-related AI models. However, there are additional challenges in the process of collecting child speech data [43], explaining the limited number of child-speech datasets available for research purposes.

A. EXISTING CHILD SPEECH DATASETS – DEFICIENCIES

There are some English child-speech datasets publicly available to researchers. Some of these [28], [29] were built using the approach of recruiting child speakers for recording sessions in professional recording studios, while others, for example, the MyST dataset [30] were built using a tablet or smartphone based app to record children's conversational speech remotely. For the latter, audio quality is highly dependent on the consumer device that the app runs on. All of these datasets feature several drawbacks, which affect data quality and introduces challenges to the use of said data in training speech-related AI models such as ASR and TTS. Invariably, major cleaning, filtering, annotation and other pre-processing of the data becomes necessary. A summary of the statistics and pros and cons of these child-speech datasets are presented in Table 1, along with some adult speech datasets for comparison.

A common problem with many of the child speech datasets is that they are relatively small/short in duration, as can be seen in Table 1, and are simply not enough in terms of duration (hours) to train a speech model on their own. Another problem is the poor quality of recorded speech samples. Some datasets are generally of poor quality due to the recording devices and/or environments used to while capturing the data; for examples audio samples may have too much background noise, noise from recording gear, or very low gain. Lastly, some datasets have several bad speech samples.

For instance, in Table 1, MyST is the largest child dataset and has a lot of data (approx. 393 hours) from multiple speakers; but many of the utterances are too short or too long, non-meaningful or indiscernible and noisy. In addition, much of the dataset is not annotated, or annotations are of poor quality and cannot be used for training speech models [24].

B. CHALLENGES IN BUILDING CHILD-SPEECH DATASETS

Building a clean speech dataset even for adults is not an easy task. It requires a specially prepared environment (recording studio), the right recording and storage devices, as well as recruitment of speakers. Child speech data can also be collected using this traditional method of recruiting speaking actors for recording sessions in media studios; however, in the case of children, additional difficulties are introduced.

- Recruitment and data protection: The processes of recruiting child speakers (actors) and complying with data protection laws can be both expensive and time-consuming and must involve the parents or legal guardians of the children, as children cannot give their own legal consent.
- Low concentration and short attention spans: children have relatively lower levels of focus and shorter spans of attention, which could cut recording sessions short.
- Poor acoustic and linguistic capabilities of the youngest group of children.
- Poor quality of recording devices and environment.

Another approach that can be used to gather children's speech involves collecting audio recordings from the Internet, for example from YouTube or through a dedicated recording application. With this approach, a different set of challenges are faced:

- Limited number of videos with children as main actors.
- Short video/utterance durations.
- Background noise and music
- Lack of transcriptions and annotations.

C. RATIONALE FOR THIS RESEARCH

Taking all the above challenges into consideration, there is a need for alternative ways to build larger child speech datasets to facilitate the development of child-friendly speech technologies. To this end, the goal of this study is to explore the potential of augmenting adult speech to provide additional child-like speech samples to complement existing child-speech datasets. The resulting synthetic child voices can be used to generate more synthetic child speech with the appropriate (child-like) linguistic content using a fine-tuned TTS model.

TABLE 1. Summary of child speech research datasets with statistics & pros and cons.

Dataset	Main Statistics			Pros	Cons
	Type	Duration (hrs)	No. Speakers		
My Science Tutor (MyST) [30]	Child (grades 3-5)	393	1371	Large amount of data	Noisy (both audio & transcripts) Not fully transcribed
PF-STAR [27] (German, Italian, English Swedish)	Child	Approx. 65	611	Clean Fully transcribed	Small in size
CMU Kids [29]	Child (Ages 6-11)	Approx. 8.9	76	Clean Fully transcribed	Small in size
CSLU [28]	Child (grades 0-10)	100	1178	Transcribed (Scripted and spontaneous)	Noisy Extremely short utterances (single words)
LJ Speech [31]	Adult (female)	25	1	Clean, fully transcribed	Small in size
Librispeech [32]	Adult (male & female)	982	2,484	Large amount of data, fully transcribed	-
LibriTTS [37]	Adult (male & female)	586	2,456	Large amount of data, clean, fully transcribed, TTS-ready	-
VCTK [38]	Adult (male & female)	44	109	Clean, fully transcribed	Small in size

D. RELATED WORKS

To improve the performance of ASR models for children's speech, some researchers have adopted similar data augmentation techniques. For example, Shahnawazuddin et al. [44] proposed a prosody modification (i.e., pitch and speaking rate scaling) using a Zero-Frequency Filtering based Glottal Closure Instants (ZFF-GCI) anchoring approach. The authors used these modifications to introduce more variability in order to achieve speaker independent ASR and reported improvements in accuracy over their baseline for both adult and child test sets (ASR).

Bhardwaj et al. [45] also used pitch and speaking rate modification to improve performance of Punjabi ASR system on children's speech. It uses the ZFF-GCI method for Linear Prediction based Pitch Synchronous Overlap and Add (LP-PSOLA) together with speaker adaptive training and achieves an improvement in recognition rate for Punjabi child speech. Chen et al. [46] applied multiple modifications including pitch, tempo, speed, and volume perturbations to both adult and child training datasets to diversify and increase the amount of available training data to improve child ASR.

The idea of generating synthetic child-like speech from adult speech was explored by Singh et al. [47]. In their work, they applied spectral modifications, namely Linear Predicting Coding (LPC)-based segmental warping perturbations (LPC-SWP) and formant energy perturbations (FEP), to adult data to generate child-like speech for data augmentation, and demonstrated an improvement in WER on both children and adult test sets when these modifications were combined with vocal tract length perturbation (VTLP).

Most of these works used different algorithmic approaches to apply prosody-based modifications (pitch and speaking rate scaling) to the speech, and the modifications were applied in a somewhat randomized manner. That is, both increasing and decreasing adjustments were applied to the audio features (e.g., pitch and speaking rate). In addition, the quality of the modified speech generated was not assessed in detail.

In this work, the goal is to generate/create synthetic child-like speech data, and we consider augmenting the pitch and speaking rate of adult speech to achieve this using a publicly available phase-vocoder based sound manipulation tool. To determine the timestamps of words and spaces where the speaking rate should be reduced, a forced alignment system based on an ASR model is used. In addition, we employ a speaker encoder model to visualize and compare the adults' and children's speaker embeddings in a common latent space before and after modifications. The contributions of this paper are as follows: a) exploring an alternative algorithm approach for the modification of adult speech (to make them more child-like through pitch and speaking rate adjustments),

b) conducting Mean Opinion Score (MOS) studies to provide a qualitative evaluation of the augmented/modified speech, c) scaling the augmentation to generate large amounts of synthetic child-like speech, d) conducting a proof-of-concept ASR experiment (example application) to provide a quantitative evaluation of the augmented adult speech.

The rest of this paper is organized as follows: Section II presents foundation technologies used in this research. Section III describes the methodology and Section IV presents the experiments conducted. Results and discussions are presented in Section V. Section VI shows an example application.and finally, Section VII presents our conclusions and future work.

II. FOUNDATION TOOLS AND TECHNOLOGIES

To develop our augmentation pipeline, we need to use a number of specialized tools to modify the pitch and control the duration of speech samples. In this section we introduce these tools, outline their features and discuss their role in the pipeline.

Different tools were considered for the tasks defined. The Combinatorial Expressive Speech Engine (CLEESE) [48] was selected to implement these augmentations because it offers a combination of ease-of-use and flexibility by allowing transformations to be applied to specific segments of the input speech sample where desired.

A. THE COMBINATORIAL EXPRESSIVE SPEECH ENGINE (CLEESE)

CLEESE is a python toolkit that can be used to perform deterministic or random transformations on input sound. Several features of the input sound can be modified, including the pitch, duration, and gain (amplitude). Originally designed to generate many random variations of a single input sound, CLEESE can also be used to perform individual and user-determined transformations, and the transformations can be either static or time-varying [48].

Using the phase-vocoder digital audio technique, CLEESE first takes the Short-Time Fourier Transform (STFT) of audio files, which decomposes each frame (segment) of the audio file into its frequency coefficients. Then CLEESE modifies the frames' STFT coefficients as required. For example, it shifts a frame's frequency coefficients to higher frequency positions to achieve a higher pitch [48]. After applying the modifications, CLEESE then generates a modified time-domain signal from the manipulated frames by applying a variety of techniques to ensure continuity or phase-coherence of the resulting sinusoidal components [48].

CLEESE operates by passing user-defined or random breakpoint functions (BPFs) to a spectral processing engine together with other parameters for processing of the sound. The BPFs are functions that determine how transformations vary over the duration of the sound, in other words, they define one or more segments (time-windows) of the input sound where specified modifications should be applied. For each BPF, a transformed version of the input sound is generated.

For the pitch, time and gain transformations, the BPFs are temporal and are specified as two-column matrices. Each row (breakpoint) in a BPF matrix has two elements: time and value. The time indicates where the next modification should begin from, and the value indicates the amount of modification to be applied. The desired transformation is specified separately in a configuration file. With the specified transformation, CLEESE modifies the input sound along the corresponding dimension (pitch, time, or amplitude) while maintaining the other dimensions constant. CLEESE can also perform chained transformations; for example, apply pitch shifting followed by time shifting.

B. CLEESE TRANSFORMATIONS

1) PITCH-SHIFT TRANSFORMATION

Pitch-shifting involves shifting or displacing the fundamental frequency in a given audio frame to a different (higher/lower) frequency. In this study, the fundamental frequencies are shifted to higher frequency points specified in the BPFs along with the corresponding times where the modifications should start. To determine the new frequency point, CLEESE takes a pitch-shift factor, a value expressed in units of cents (a cent is one hundredth of a semitone), provided in the BPF and uses it to compute the new frequency with respect to the original frequency. As an example, to shift the pitch of the input audio by 2 semitones, a pitch-shift factor of 200 cents is provided in the BPF. Pitch-shift factors less than 0 cents correspond to lowering the pitch, factors greater than 0 cents correspond to raising the pitch, and a factor of 0 cents implies no change or shift in pitch [48].

2) TIME-STRETCH TRANSFORMATION

The time-stretching transformation involves shifting the audio frames from their original positions to earlier or later points. Similarly, for the time-stretching, CLEESE takes a time-shift factor from the given BPF and uses that to determine the new position of a frame. A time-shift factor less than 1 corresponds to compressing the sound, a factor greater than 1 corresponds to stretching the sound, and a factor of 1 implies no change in the original sound duration [48]. For example, using a time-shift factor of 2 doubles the duration of the audio, i.e., a 3-second-long audio will become 6-seconds-long after modification, if the modification is applied to the full length of the input audio.

C. WAV2VEC2 FORCED ALIGNMENT SYSTEM

The wav2vec2.0 forced alignment system[1,2] uses the wav2vec2.0 [4] ASR model for extracting acoustic features from the audio and estimating the frame-wise label probabilities. It then constructs a Trellis matrix using the ground-truth

[1]https://github.com/pytorch/audio/blob/main/examples/tutorials/forced_alignment_tutorial.py

[2]https://pytorch.org/audio/stable/tutorials/forced_alignment_tutorial.html

TABLE 2. The wav2vec2.0 alignment system outputs all words in an utterance, their respective start and stop times, as well as the confidence score for each alignment.

Confidence _level	Word_label	Start_time	Stop_time
0.53	SHE	0.604	0.725
0.80	HAD	0.765	0.926
1.00	A	0.967	0.987
0.84	THIN	1.108	1.430
0.80	AWKWARD	1.792	2.175
0.91	FIGURE	2.236	2.538

transcript of the utterance, which shows the probability of the transcript's labels at each timestep. The system then finds the most likely path through the Trellis matrix, producing the alignments between the ground-truth transcript's words and the spoken audio. The output of the forced alignment process is the start and end timestamps for all words in an utterance as shown in table 2.

D. SPEAKER EMBEDDINGS

A speaker embedding is simply a representation of a speaker's identity in the form of a fixed size vector given an utterance, and regardless of the utterance duration. Speaker embeddings can be plotted in an embedding vector space to visualize how multiple speakers relate to each other. Speaker embeddings are commonly used for speaker recognition tasks [16], [49] and more recently, to improve multi-speaker TTS models [8]. In addition to the speaker identities, speaker embeddings may carry information about other paralinguistic information such as prosody or emotion and gender of a speaker.

Different approaches have been proposed to encode speaker embeddings, and these include identity vectors (i-vectors) [50], which are low-dimensional projections of the differences between a speaker's pronunciations and the respective overall average pronunciations; (d-vectors) [51], which are deep neural network (DNN) based and extracted from a hidden layer of a model trained to predict speaker identities; and x-vectors [52], which are also DNN based but capture segment/utterance level information as well as frame-level information by using either statistical or max-pooling method to gather the frame level information as segment level representation [53].

III. METHODOLOGY

In this section, we describe the implementation of the proposed adult-to-child speech augmentation process. The python toolkit, CLEESE, is used to perform two key transformations to the adult speech data with the aim of transforming them to child-like speech. Fig. 1 shows a flow diagram of the overall augmentation process.

First, we triage the adult speakers by comparing the cosine similarities of their speaker embeddings to child speaker embedding prior to the augmentation process, see Fig. 2. This is done by computing the mean child speaker embedding as well as the mean speaker embedding per adult speaker. Each adult speaker's cosine similarity to the mean child embedding is computed, and the value is compared to a threshold value for a selection decision to be made. More details on this in section VI, B.

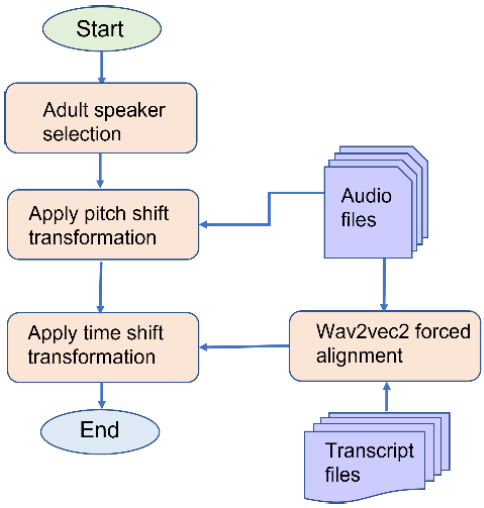


FIGURE 1. Flow diagram for the adult-child speech augmentation process.

Next, we apply the pitch-shifting transformation to the utterances of the selected speakers. For each utterance, the pitch transformation is applied to the full utterance length. To achieve this, a BPF is created with one break point (time: start of utterance and value: the desired pitch-shift factor e.g. 100 cents i.e., 1 semitone). CLEESE applies the transformation from the specified time stamp to the end of the utterance unless another breakpoint is encountered. Therefore, a single breakpoint (row) in the BPF modifies the full utterance.

Next, the time-stretching transformation is applied to the pitch-shifted utterances. To successfully stretch the desired segments of the sound, the exact start and stop times for the segments are needed to create the appropriate BPFs for stretching. For this, the wav2vec2.0 based forced alignment system is employed to align the adult speech with their corresponding transcripts. Based on the word timestamps, the start and end times of all "white spaces" in the utterance are derived and used in creating BPFs for the time-stretching transformation. The start time for each word and white space is used as a breakpoint in the BPFs, and different stretch factors are used for words vs whitespaces.

IV. EXPERIMENTS

The proposed techniques for augmentation were implemented on an NVIDIA GeForce RTX 2080 Super GPU, and to scale our experiments we used an NVIDIA RTX A6000 GPU.

A. PRELIMINARY TESTS

Pitch-shift and time-stretch transformations were applied to randomly selected subsets of two adult speech datasets: LJ

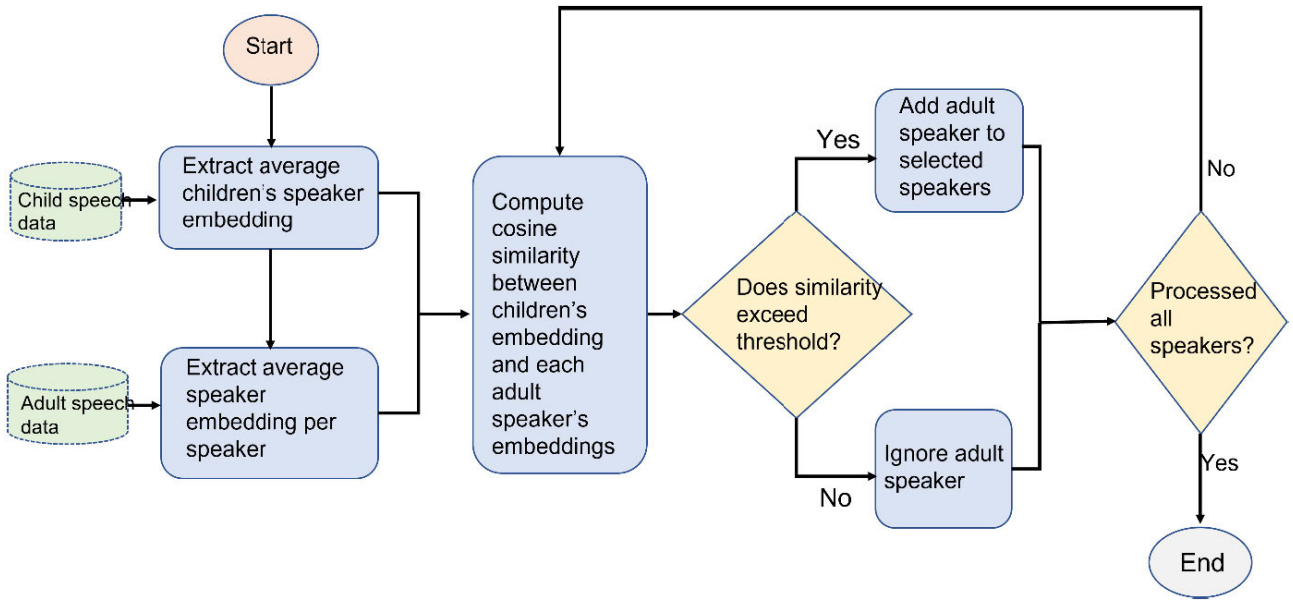


FIGURE 2. A flow diagram for the adult speaker selection process.

speech [31] and voxceleb1 [34]. Pitch-shift factors in the range of 100 cents (1 semitone) to 800 cents (8 semitones) were tested on both male and female speakers, and random time-shift factors in the range of 2 to 4 were also tested. The goal was to determine approximately the range of pitch-shift and time-shift factors that will make sense to use in future experiments.

The time-stretched utterances were qualitatively evaluated by listening to them, and it was observed that a time-shift factor of 4, which quadruples the audio length, resulted in extremely sluggish augmented utterances, and even a factor of 2, which doubles audio length, resulted in utterances that were still a bit too slow. Another observation made was that stretching the individual words in the utterance made them sound unrealistic.

For the pitch-shift transformation, we observed that, with the range of values (pitch-shift factors) that achieved desired results on some of the speaker identities, other speaker identities did not sound realistic, even after extending the range of pitch-shift factors. From these initial tests we determined that not all adult voices can be successfully tuned to sound child-like.

To resolve this and allow a larger study to be conducted, it was necessary to first triage and determine the adult speakers whose voices are more suitable for transforming into natural child voices. This could be achieved by projecting both adults' and children's speaker embeddings into a latent speaker embedding space for comparison.

B. INITIAL EXPERIMENTS

1) COMPARISON OF ADULTS' AND CHILDREN's SPEAKER EMBEDDINGS

To compare the adults' and children's speaker identities, a Generalized End-to-End (GE2E) Loss [49] based speaker embedding (encoder) tool known as Resemblyzer [54] was used. It uses a d-vectors based speaker encoder model [49] which uses the GE2E loss for optimization. It also has multiple functionalities for visualizing and comparing the extracted embeddings using the Unified Manifold Approximation and Projection (UMAP) for dimension reduction.

Initially, the speaker embeddings of multiple speakers (both adult and children) were plotted via UMAP with the aim of finding adult speaker embeddings closest to the children's speaker embeddings. In Fig. 3, we show some speaker embeddings in a UMAP plot for visualization.

All male speaker embeddings are marked with black crosses, female speaker embeddings are marked with blue triangles and all child speaker embeddings are marked with red circles. The children's embeddings cluster in a small section of the embedding space. However, it was challenging to accurately identify the adult speakers that are closest or most similar to children by visual inspection. Therefore, it was decided to perform a cosine similarity-based comparison and select speakers with the highest similarity values for the main augmentation experiments.

2) COMPARISON OF EMBEDDINGS BASED ON COSINE SIMILARITY

The cosine similarity score is a number between 0 and 1. A similarity of 1 means the two embeddings compared are identical, and a similarity of 0 means they are completely different. Firstly, we extracted the speaker embeddings for multiple child speakers taken from the CMU kids corpus [29]. From previous research [24] as well as initial experiments (see Fig. 3), it is known that children's speaker embeddings form a small cluster in the speaker embedding latent space;

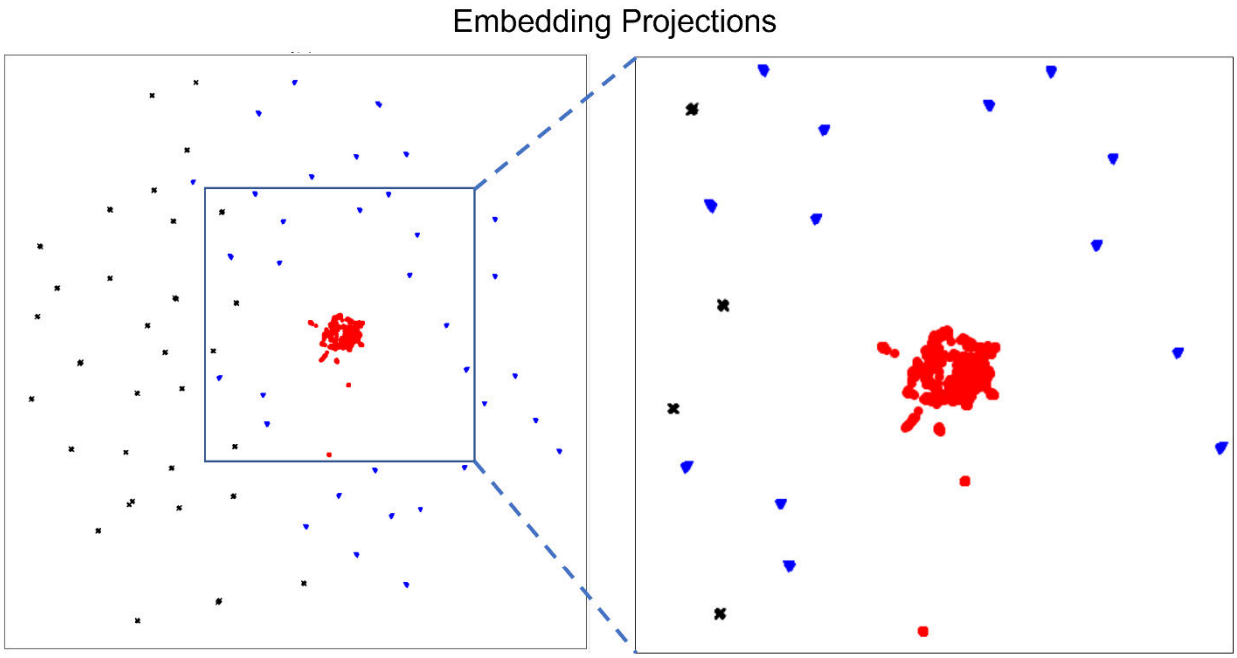


FIGURE 3. Projection of 65 Adult speaker embeddings from Librispeech: 31 male (black), 34 females (blue) and 31 child speaker embeddings from CMU Kids.

hence, we computed the mean child speaker embedding to represent the children embedding cluster.

Next, we took a random subset from the train-clean-100 subset of the Librispeech dataset [32] and computed an average speaker embedding for each adult speaker, by averaging the embeddings of their individual utterances. Then we compared them to the mean child speaker embedding using the cosine similarity metric. A flow diagram showing the flow of this process is seen in Fig. 2.

All adult speakers whose embeddings exceeded a predefined threshold of 0.65 were selected for augmentation as in equation 1. This threshold was chosen by listening to some of the utterances and observing their corresponding similarities. Fig. 4 shows some examples of the cosine similarities computed. More statistics regarding the cosine similarities are shown in the next section.

$$Dec = \begin{cases} 0, & sim\_score < 0.65 \\ 1, & sim\_score \geq 0.65 \end{cases} \quad (1)$$

where Dec is adult speaker selection decision, sim_score is the computed cosine similarity score between an adult's speaker embedding and the average child speaker embedding.

3) AUGMENTATION PROCESS

Further tests were done on the selected speakers (i.e., adult speakers whose cosine similarities exceeded the threshold) thereafter. Two separate ranges of pitch-shift factors were empirically chosen for the two genders. This was done by listening to the pitch-shifted utterances and rating them in terms of how convincingly child-like they sounded. For male

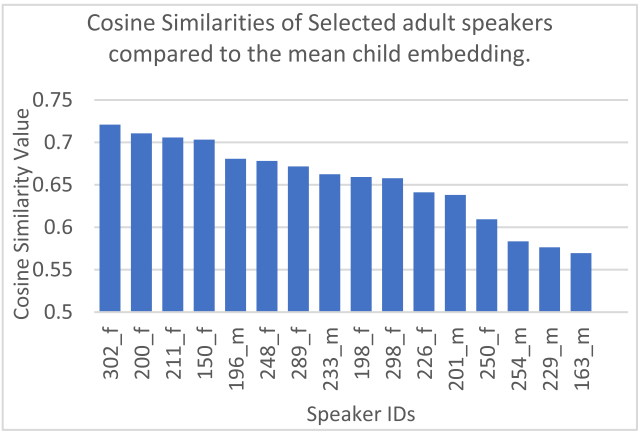


FIGURE 4. Original cosine similarities showing how similar Librispeech adult speakers are to mean CMU kids child speaker embedding. The speaker gender is suffixed to the Speaker IDs.

and female speakers, the ranges of 500 to 700 cents and 100 to 300 cents were chosen respectively.

Based on the observations made about the time-stretched utterances in the preliminary tests, it was decided to stretch only the pauses (whitespaces) between all words in the utterances, as well as the unusually long words, without stretching every single word. For stretching all the whitespaces, we first used a time-stretch factor of 2 (i.e., we doubled the length/duration of pauses) and then reduced it to a time-stretch factor of 1.8 after qualitatively evaluating a few of the augmented utterances. In addition, we identified the unusually longer words in the utterances - that might be

TABLE 3. Number of librispeech train-clean-100 speakers above and below the cosine similarity threshold.

Speakers	Total No.	No. speakers above threshold	No. speakers below threshold
All	251	121	130
Males	126	27	99
Females	125	94	31

TABLE 4. Final shift factors used for pitch and time transformations.

Transformation	Shift-factors
Pitch-shift (female)	100, 200, 250, 300, 350, 400 (cents)
Pitch-shift (male)	500, 600, 700 (cents)
Time-stretch	1.8 (for white spaces), 2.0 (for longer words)

difficult for children to pronounce - and stretched them using a factor of 2. This was done by computing the duration of each word and comparing it to an empirically chosen word length threshold.

C. MAIN EXPERIMENTS

For the main experiments, we used all the data in the train-clean-100 subset of Librispeech [32] as the adult speech dataset. A subset of the CMU kids dataset [29] was used as the child speaker set, specifically the Fort Pitt (FP) subset. Firstly, the adult speakers most proximate/similar to children in terms of speech/voice were determined by performing the cosine similarity comparison explained in Section III, B using the same decision threshold value of 0.65 as in the initial experiment. Table 3 shows the number of adult speakers above and below the cosine similarity threshold.

Once the most similar speakers were selected, the two augmentation techniques explained in Section II, B; namely, pitch-shifting and time-stretching transformations were applied to all individual utterances of the selected speakers. The same pitch-shift and time-stretch factors chosen in the initial experiment were applied here. First, we applied the pitch-shifting transformations and then applied the time-stretching transformation on the output of the pitch-shifting transformation. Table 4 shows the final shift factors used for the pitch-shift and time-stretch transformations.

This resulted in multiple sets/folders of data per speaker, each containing utterances augmented with different augmentation parameters. Specifically, the sets of utterances differed in terms of pitch-shift factors only, as the time-stretching parameters were kept constant for all sets and all genders.

D. OBJECTIVE EVALUATION

In the initial experiment section, the cosine similarity value served as a good metric to determine the proximity of adult speaker embeddings to the average child speaker embedding. Therefore, to objectively evaluate the augmented speech, it made sense to recompute the cosine similarities between adult speakers' average embedding (after augmentation) and the average child embedding. After recomputing the cosine similarities, we observed that there was a general increase in the similarity values for all the speakers. Fig. 5 below shows the cosine similarities of selected speakers before and after transformations were applied. A similarity score of 1 would indicate that a speaker is exactly the same as the average child speaker. Table 5 also shows statistical analysis of the adult speech data before and after augmentation. Note that the cosine similarities of all individual child speakers' embeddings to the mean child embedding were in the range of 0.9 to 0.973, except one child (0.837).

TABLE 5. Statistics of cosine similarities for librispeech train-clean-100 before and after augmentations.

Data	Before augmentation		After augmentation	
	Mean	STD	Mean	STD
All above threshold (121)	0.69	0.032	0.83	0.034
20 spkrs in MOS study	0.740	0.022	0.86	0.027
16 spkrs in MOS	0.745	0.019	0.87	0.018

E. SUBJECTIVE EVALUATIONS

While the increase in the cosine similarity of an augmented adult speaker gives a strong indication that the augmentation pipeline is achieving its primary goal, it is not possible to judge how realistic or intelligible the augmented voice is. In the case of some subjects, it was noted that while the cosine similarity was high, the corresponding speech was occasionally distorted and unrealistic.

For this reason, it was decided to conduct a human listener evaluation study to validate how realistic the augmented speech from a speaker is and confirm that it remains intelligible. Such a study can also help confirm the best speakers and the optimal augmentation parameters to use for individual speakers to build a larger augmented speech dataset – a core goal of this research.

To subjectively evaluate the augmented speech samples, the MOS [55] subjective evaluation method was applied. MOS evaluation is widely used to evaluate speech models, such as TTS and Voice Conversion (VC) models, by asking human evaluators to rate various aspects of speech quality such as naturalness, intelligibility, similarity, etc.

1) DESIGN OF MOS STUDY

There were three specific goals for the study: i) Determine the optimal pitch-shift factor per speaker, ii) Determine how realistic (convincingly child-like) the augmented utterances sound and iii) Determine whether the augmented utterances are distorted beyond understanding or if they remain intelligible. To achieve the goals of the study, three questions that

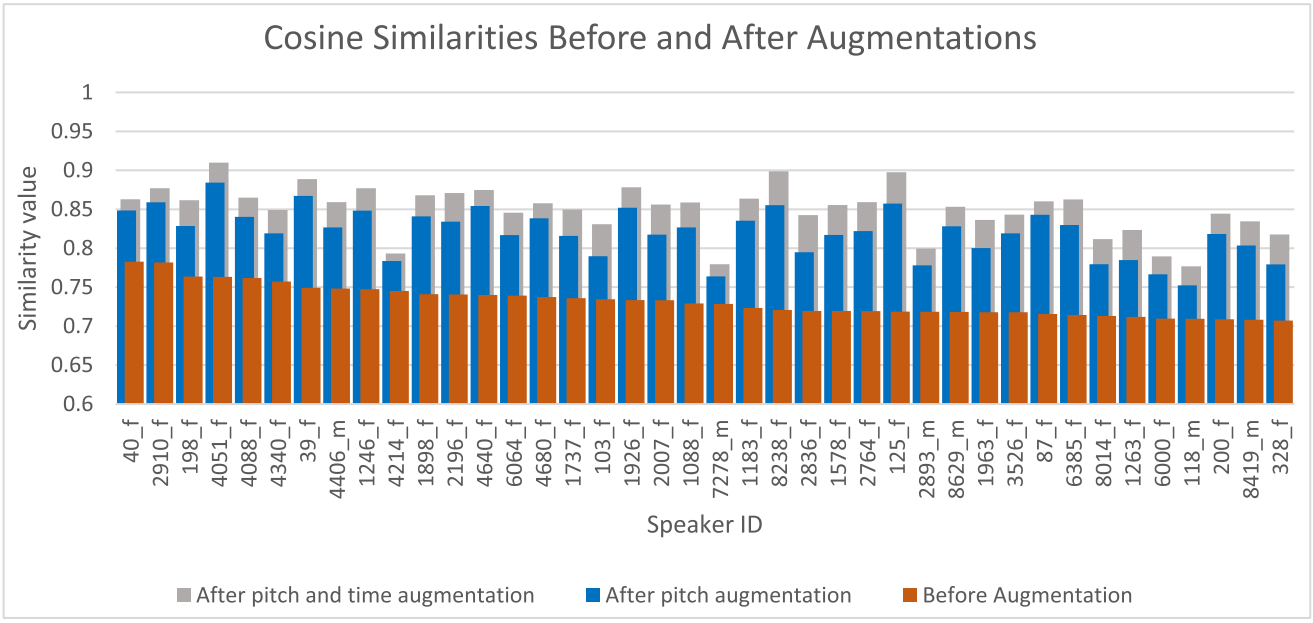


FIGURE 5. Increases in cosine similarity between adult and child speaker embeddings after pitch shifting and time stretching.

capture the information required were chosen and presented to the evaluators.

For the first goal, evaluators were provided with multiple variations per utterance and asked to select the most child-like sounding one. The difference between the variations are the pitch-shift factors used in the pitch-shift transformation. The sample selected for the first question is used in the remaining questions. Secondly, evaluators were asked to rate the selected sample in terms of how convincingly child-like or how realistic it sounds on a scale of 1 to 5. Note that the linguistic contents of the utterances are adult-like and very different from the typical linguistic content of child speech. Evaluators were given prior notice and were asked to disregard the adult-like linguistic content while rating the convincingness.

Thirdly, evaluators were asked to rate the same selected sample in terms of intelligibility on a scale of 1 to 5. Evaluators were restricted to only 5 grading points (i.e., 1, 2, 3, 4 and 5); they were not allowed to give intermediate scores, such as 2.5. Evaluators were also asked to identify the gender of the speaker by choosing one of three options: Boy, Girl and Can't say. Finally, evaluators were also given the option to leave comments if they had any. Table 6 shows explanations of the scales for convincingness (question 2) and intelligibility (question 3) following approach in [24].

With this design, a first MOS study (Study A) was conducted on utterances augmented with the following pitch-shift factors: 100 cents, 200 cents and 300 cents, meaning for each utterance, there were three variations for evaluators to choose from. After this first study was completed and the results were processed, it was decided to conduct a second MOS study (Study B) to refine the outcome of the first. In particular, we wanted to see the effect of including utterances augmented with higher and more finely granulated pitch-shift factors as compared to the first study: 250, 300, 350 and 400 cents. This is because in the first study, the same variation of utterance (augmented with the highest pitch-shift factor of 300 cents) was selected by almost every evaluator as the most child-like for all female speakers, so the range of pitch-shift factors to investigate clearly needed expansion.

Study A:

In the first evaluation, there was a total of 30 evaluators, mainly drawn from an undergraduate engineering class. These were divided into two groups of 15 evaluators. Augmented speech samples were taken from Librispeech speakers. 20 speakers were chosen for the MOS study, after listening to samples from all their recording sessions to check for noise and rate the quality. This was done after triaging the adult speakers as described in Section III B. They included 16 female speakers and 4 male speakers with the highest cosine similarities and high-quality audio samples. Each group of 15 evaluators was given a unique set of 10 different speakers to review (8 females and 2 males). The purpose was to reduce the total number utterances per evaluator. To diversify the phrases, each evaluator group was further divided into 3 subgroups and each subgroup was given a unique (randomly selected) set of 2 phrases per speaker, resulting in a total of 20 phrases to evaluate per evaluator. They were given three augmented samples (variations) per phrase: A, B and C corresponding to pitch shifting factors of 100, 200 and 300 cents, respectively.

Study B:

In the second evaluation study, augmented samples from only the 16 female speakers out of the top 20 speakers

TABLE 6. Explanation of rating scales for question 2 (convincingness) and 3 (intelligibility).

Rating (Score)	Question 2: Convincingness	Question 3: Intelligibility
1	Unconvincing	Bad voice intelligibility and bad word comprehensibility; None/only a few words are intelligible.
2	Slightly convincing, not very like a normal child's voice	Weak voice intelligibility and weak word comprehensibility; more than 50% of the words unintelligible
3	Plausible but not very convincing as a normal child's voice	Weak voice intelligibility but plausible word comprehensibility; more than 50% of words are intelligible but require significant concentration.
4	Quite convincing, sounds very close to a child's voice	Mediocre but plausible intelligibility and comprehensibility; most words intelligible and relatively easy to identify.
5	Very convincing, sounds exactly like a child's voice	Good intelligibility and comprehensibility; All words are quite clear and comprehensible.

TABLE 7. Summary of the data distribution for mos studies a and b.

	Study A		Study B	
Total no. of Evaluators	30		60	
Subgroups	I	II	I	II
No. of Evaluators	15	15	30	30
No. of Male IDs	2	2	-	-
No. of Female IDs	8	8	8	8

(the same speakers as in the first study) were evaluated. There were 60 evaluators, again mostly engineering students, divided into 2 main groups of 30 evaluators. Each group was further divided into 3 subgroups of 10 students, similar to the approach used in Study A. This time, each evaluator received 16 phrases from 8 speakers: two phrases per speaker as in the first evaluation. Specifically, there were four variations per utterance/phrase: A, B, C and D corresponding to pitch shifting factors of 250, 300, 350 and 400 cents, respectively. Information about evaluators for the two MOS studies is presented in Table 7.

V. RESULTS AND DISCUSSION

In Section III, we described our adult-to-child speech augmentation experiments using the two augmentation techniques described in Section II, with a goal of making the adult voices sound child-like. We also conducted two MOS studies to evaluate the quality of the synthetic child-like speech. In this section, we present and discuss the results of our experiments.

Tables 8 and 9 show the results obtained from the first and second MOS studies, respectively. More detailed presentations of the MOS evaluation results are shown in Tables 12 and 13 in the Appendix. The results of the subjective evaluation showed that the utterances of adult female speakers consistently ranked with higher scores for intelligibility and significantly higher scores for convincingness. We had anticipated this result as only 4 males ranked in the top 20 speakers from Librispeech train-clean-100. It is clear that female speakers offer a better starting point to build synthetic child voices than male speakers.

As shown in both Table 12 and Table 13, the optimal pitch-shifting factors for the female speakers lie in the range of 300 to 400 cents. Augmenting the pitch above this range causes the augmented speech to sound more chipmunk-like rather than child-like. For the male speakers, the pitch-shift factor of 600 cents was selected for 3 out of 4 speakers but the augmented speech were unconvincing as child voices, with a very low average MOS score of 1.76.

TABLE 8. Mean and standard deviation (std) of convincingness and intelligibility MOS scores (C-MOS and I-MOS) from Study A.

Speakers	No.	C-MOS (STD)	I-MOS (STD)
Female	16	3.37 (0.37)	4.32 (0.20)
Male	4	1.76 (0.37)	3.87 (0.39)
All	20	3.05 (0.75)	4.23 (0.30)

TABLE 9. Mean and standard deviation (std) of convincingness and intelligibility MOS Scores (C-MOS and I-MOS) from Study B.

Speakers	No.	C-MOS (STD)	I-MOS (STD)
Females	16	3.70 (0.35)	4.36 (0.25)

TABLE 10. Details of the synthetic and original data used in finetuning.

Finetuning data	Adult/child	Orig/Synt	Duration(hrs)
Original_12h	Adult	Orig	12
Augmented_17h	Child	Synt	17
Original_220h	Adult	Orig	220
Augmented_311h	Child	Synt	311
MyST_55h	Child	Orig	55

TABLE 11. WER of ASR models finetuned with synthetic and original speech data.

Model	Group	Pretraining	Finetuning dataset	WER MyST_10h	WER PFS_10h	WER CMU_9h	WER Devclean_9h
1	Group A	Librispeech	Original_12h	19.95	25.10	18.95	5.78
2		Librispeech	Augmented_17h	20.11	20.48	19.14	6.58
3		Librispeech	Original_12h + Augmented_17h	18.17	18.11	16.07	5.49
4	Group B	Librispeech	MyST_55h	8.13	17.67	16.47	7.72
5		Librispeech	MyST_55+Original_12h	8.10	16.76	15.45	5.62
6		Librispeech	MyST_55h + Augmented_17h	7.98	14.015	15.02	4.87
7		Librispeech	MyST_55h +Original_12h+Augmented_17h	7.95	14.85	13.92	5.54
8	Group C	Librispeech	Original_220h	15.09	16.59	14.41	4.39
9		Librispeech	Augmented_311h	17.42	15.86	15.09	4.83

The overall C-MOS score of the most child-like samples was approximately 3.0 when adult male speakers were considered, and 3.7 when only adult female speakers were evaluated (study B). Both convincingness MOS values are above average and implies that the augmented samples are reasonably convincing in terms of human perception and very convincing when only female speakers are used in the study.

A relatively higher I-MOS was obtained for the augmented samples of both genders, showing that generating synthetic child voices using our proposed method does not significantly degrade the intelligibility of the original speech samples.

Note that there are limitations in going from adult speech to child speech; for example, the linguistic content of adult speech data is completely different from the typical linguistic content of children's speech. For this reason, tuning the pitch and speaking rate of adult speech would not make the speech sound completely natural as child speech in terms of the linguistic content. However, these tunings can make the voices alone sound reasonably child-like, which is the target for the current study.

The mean cosine similarity of adult speakers after augmentation was 0.83 for all speakers exceeding the similarity threshold and 0.87 for the top 16 female speakers (see Table 4), whereas the mean cosine similarity of the individual child speakers was 0.94, indicating that there is still potential to further augment the adult speakers to sound closer to child speakers. This suggests that additional prosodic features and paralinguistic elements could be investigated and added into our augmentation strategy to improve the cosine similarity score of the adult speakers.

Finally, to validate the augmented child speech data in a practical application, we next run some ASR fine-tuning experiments, as presented in the next section.

VI. VALIDATION OF THE AUGMENTED SPEECH: EXAMPLE APPLICATION – ASR FINETUNING

In this section, as an example application, we conduct semi-supervised ASR finetuning experiments with our augmented adult speech dataset, to show that the augmented speech could achieve improvement over simply using additional adult speech to finetune the ASR for child speech.

Note that the main goal of our study was to explore data augmentations to make adult speech data sound more child-like (i.e. closer to child speech data) in order to provide more child-like data for training, testing and validation of ASR and TTS models to improve their performance on real child speech. Here, we show that finetuning a semi-supervised ASR model with augmented adult speech data can improve the ASR model's performance on child speech. We show that even when finetuned with adult-only speech data, the performance of the model improves to an extent; however, there is some additional improvement when the augmented adult speech is used.

We used the state-of-the-art (SOTA) wav2vec2.0 ASR model [3], which uses a self-supervised learning approach and has a two-step training process. First, the model is pretrained on a large amount on unlabeled speech, then it is finetuned on labelled speech data for a downstream task, such as ASR. We used a publicly available pretrained wav2vec2.0 model, which was trained on approximately 1000 hours of unlabeled Librispeech data [32]. This model was then finetuned with different combinations of our augmented datasets in the various finetuning experiments as presented in the next sub-section. The aim was to compare the performance of an ASR model finetuned with real child and/or adult speech versus the same model finetuned with our augmented data (synthetic child-like speech). The Word Error Rate (WER) metric was used to measure the performance of the finetuned ASR models.

A. ASR FINETUNING DATASETS

We created two sets of synthetic child speech:

- *Augmented_17h:* Contains augmented utterances from the 16 female speakers of the train-clean-100 Librispeech dataset, whose speaker embeddings are most similar to an average child embedding from the

CMU-kids corpus by cosine similarity. The female speakers were selected by ranking all female speakers by their similarity score. This data totals approximately 17 hours in duration.

- *Augmented_311h:* Contains augmented utterances of all female speakers in Librispeech train-clean-360, train-clean-100, dev and test sets combined, whose similarity score to the average real child embedding from the CMU-kids corpus is above 0.6. This data totals approximately 311 hours in duration.

We also used original (non-augmented) adult speech from Librispeech [32] and real child speech data from the MyST child speech corpus [30] for our finetuning experiments:

- *Original_12h:* Contains 12 hours of original adult speech.
- *Original_220h:* Contains 220 hours of original adult speech.
- *MyST_55h:* Contains 55 hours of cleaned MyST child speech, which was prepared according to [56]

The Original_12h and Original_220h sets are the original Librispeech (adult speech) counterparts of the Augmented_17h and Augmented_311h sets, respectively. Note that there is an increase in the number of hours of speech data when augmenting from Original_12h to Augmented_17h and from Original_220h to Augmented_311h. More information about the finetuning datasets can be found in Table 10.

B. ASR FINETUNING EXPERIMENTS

To test our hypothesis of a lower WER on child test data after finetuning on our synthetic child-like speech data, we prepared multiple finetuning experiments. The details of these experiments are presented in Table 11. The experiments were divided into three groups- A, B and C. Group-A experiments contained only the Original and Augmented datasets. MyST_55h was added for the finetuning experiments in Group-B in addition to the Original and Augmented datasets. Group-C experiments used the combined Librispeech datasets across all speakers, both original and augmented versions. All the groups used a pretrained wav2vec2.0 model which was pretrained on 960 hours of Librispeech data.

We used four test datasets to test our finetuned models at the inference stage. These datasets were prepared in accordance with our previous research on child speech ASR [56]. Since MyST [30] is the largest child audio corpus available publicly for research use, it was used for both finetuning and inference. This was done to see the performance when finetuning and testing on similar data distributions. We used 10 hours of MyST child speech data, 10 hours of PFSTAR British English data [27], 9 hours of CMU-Kids American English child speech data [29], and 9 hours of Librispeech dev-clean data as our test datasets. Different child speech test datasets were selected specifically to check the performance of our finetuned models on datasets that have different acoustic attributes, in conjunction with adult speech also. WER values obtained on these test datasets during inference are shown in Table 11.

C. ASR FINETUNING RESULTS

Group-A: Finetuning with Augmented_17h resulted in a decrease in WER on the PFS_10h data (British English child speech), and a slight increase in WER on the other child test sets, when compared to inference with a model finetuned on its original speech counterpart (Original_12h). Furthermore, combining just 17 hours of the augmented child speech (Augmented_17h) with original adult speech (Original_12h) leads to a slight improvement in WER on all child test sets, as well as the adult speech test data.

Group-B: This group uses the cleaned MyST_55h dataset in addition to the datasets used in Group-A experiments. Using Augmented data along with MyST child speech dataset led to a decrease in WER on all the test datasets (see model 6 in Table 10).

Group-C: This group used datasets created from large-scale augmentation. There was an 18.3x increase in dataset size from Original_12h to Original_220h and from Augmented_17h to Augmented_311h, respectively. Augmentation led to a decrease in WER on PFS_10h test data, but an increase in WER for all other datasets, which is very similar to the results of Group-A experiments.

D. DISCUSSION OF RESULTS

For Group-A, the WER decreases for all the test datasets when both original and augmented adult speech datasets were used for finetuning.

With MyST data inclusion in Group-B, we see a major decrease in WER compared to Group-A results.

Furthermore, in Group-B, it can be seen that adding augmented speech along with MyST_55h (model 6) led to decrease in WER on all the test datasets compared to using only MyST child speech for finetuning (model 4) or using both MyST and original adult speech (model 5). Also, by adding both the original and augmented speech for finetuning (model 7), an increase in WER can be observed on PFS_10h and adult data, while the WER on CMU_9h is reduced.

Using Original_220h and Augmented_311h in the Group-C experiments did not lead to improvements in ASR performance when compared with Group-B results. Comparing models 2 and 9, with an 18x increase in the amount of augmented data, respectively, the WER decreased by only 3.5 points on average on child speech.

While improvements in the child ASR performance were expected, the results from the example application do not show significant improvements using just the large amount of synthetic child speech for finetuning. This could partly be attributed to a lack of natural prosody in the augmented adult data (synthetic) when compared to real child audio. Although the synthetic speech sounds reasonably child-like in

TABLE 12. Per speaker MOS Scores and best shift factors from 1ST evaluation.

Speaker ID	True Gender	Shift Factor Selection Count			Best Shift Factor	Convincingness	Intelligibility
		100 (A)	200 (B)	300 (C)			
39	Female	1	2	27	C	3.37	4.27
40	Female	3	9	18	C	3.4	4.3
103	Female	4	6	20	C	3.87	4.57
198	Female	1	3	26	C	3.2	4.07
1183	Female	0	6	24	C	3.13	3.87
1624	Male	2	18	10	B	2.13	4.13
1737	Female	1	4	25	C	3.77	4.43
1898	Female	0	4	26	C	3.13	4.27
1926	Female	1	3	26	C	3.8	4.2
2893	Male	12	9	9	A	2.03	4.27
2007	Female	0	4	22	C	3.31	4.42
2196	Female	0	0	26	C	2.42	4.31
2764	Female	0	6	20	C	3.31	4.46
2910	Female	4	5	17	C	3.46	4.04
4051	Female	0	2	24	C	3.92	4.5
4340	Female	1	8	17	C	3.46	4.31
4680	Female	3	1	22	C	3.04	4.54
1088	Female	0	5	21	C	3.31	4.54
4406	Male	9	15	2	B	1.46	3.42
8419	Male	8	11	7	B	1.42	3.65

TABLE 13. Per speaker mos scores and best shift factors from 2ND evaluation.

Speaker ID	True Gender	Shift Factor Selection Count				Best Shift Factor	Convincingness	Intelligibility
		250 (A)	300 (B)	350(C)	400 (D)			
39	Female	2	9	16	17	D	4.02	4.71
40	Female	10	12	14	8	C	3.75	4.61
103	Female	10	10	14	10	C	3.82	4.5
198	Female	2	8	16	18	D	3.82	4
1183	Female	9	14	10	11	B	3.11	3.95
1737	Female	4	16	11	13	B	3.79	4.34
1898	Female	5	11	19	9	C	3.20	4.45
1926	Female	4	8	19	13	C	4.16	4.61
2007	Female	4	12	13	11	C	3.85	4.4
2196	Female	7	8	13	12	C	3.4	4.675
2764	Female	7	12	6	15	D	3.63	4.13
2910	Female	7	8	13	12	C	3.78	3.9
4051	Female	8	9	14	9	C	3.98	4.35
4340	Female	9	12	9	10	B	4.1	4.43
4680	Female	6	9	11	14	D	3.0	4.3
1088	Female	8	10	10	12	D	3.83	4.48

terms of pitch and speaking rate, they are still lacking natural prosody characteristics such as stammering, long pauses (due to uncertainty) and other features seen in real child audio recordings. Features of natural child speech prosody could be modeled in addition to the proposed augmentation approach, which is expected to improve WER further.

VII. CONCLUSION AND FUTURE WORK

We have presented experiments exploring the possibility of generating synthetic child voices by augmenting existing adult speech datasets. Augmenting the pitch and duration of adult speech samples generally caused them to sound more child-like, however this worked better for the female adult speakers as compared to the males. This observation was further confirmed by the results of a subjective evaluation conducted using the MOS evaluation method with a total of 72 participants. The average cosine similarity of augmented adult speech is still lower than that of real child speakers, therefore more research is required to improve the similarity of augmented speech. While the improvements in the performance of finetuned ASR models on real child speech are relatively small, they provide a validation of the approach which can be further improved with a more sophisticated set of augmentations. These are planned for future work.

We have scaled this data augmentation process to provide a large amount of synthetic child speech suitable for training child-friendly TTS, VC and ASR models and the data, along with pipeline implementation code, will be made publicly available to other researchers who wish to replicate our approach.

In future experiments, we plan to investigate and apply other tuning techniques to better augment the adult male voices as well as improve the existing augmentation techniques to better suit the linguistic content on a sentence-by-sentence basis. We also plan to investigate methods that take the natural child prosody or paralinguistic feature modeling into consideration; this could contribute to further increasing the similarity of the augmented adult (synthetic child) speech to real child speech.

APPENDIX

See Tables 12 and 13.

ACKNOWLEDGMENT

The authors would like to thank Zoran Fejzo from Xperi Corporation and the rest of the team members for their helpful discussions and feedback.

REFERENCES

[1] D. Amodei et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," 2016, *arXiv:1512.02595*.

[2] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," Aug. 2015, *arXiv:1508.01211*. Accessed: Jan. 12, 2023.

[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," Apr. 2019, *arXiv:1904.05862*.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2020, pp. 12449–12460. Accessed: Jan. 12, 2023. [Online]. Available: https://proceedings.neurips.cc//paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," May 2020, *arXiv:2005.08100*. Accessed: Jan. 12, 2023.

[6] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," Apr. 2019, *arXiv:1904.08779*, doi: 10.21437/Interspeech.2019-2680.

[7] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6124–6128, doi: 10.1109/ICASSP40776.2020.9053889.

[8] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," Jan. 2018, *arXiv:1806.04558*. Accessed: Jan. 11, 2023.

[9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," Apr. 2017, *arXiv:1703.10135*. Accessed: Apr. 7, 2022.

[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," Feb. 2018, *arXiv:1712.05884*. Accessed: Jan. 12, 2023.

[11] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," Mar. 2021, *arXiv:2006.04558*. Accessed: Jun. 21, 2021.

[12] S. Beliaev and B. Ginsburg, "TalkNet2: Non-autoregressive depth-wise separable convolutional model for speech synthesis with explicit pitch and duration prediction," Jun. 2021, *arXiv:2104.08189*. Accessed: Jan. 11, 2023.

[13] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," Oct. 2020, *arXiv:2005.11129*. Accessed: Jun. 21, 2021.

[14] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," Feb. 2021, *arXiv:2006.06873*. Accessed: Jan. 12, 2023.

[15] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: An end-to-end neural speaker embedding system," May 2017, *arXiv:1705.02304*. Accessed: Nov. 3, 2021.

[16] N. R. Koluguri, J. Li, V. Lavrukhin, and B. Ginsburg, "SpeakerNet: 1D depth-wise separable convolutional network for text-independent speaker recognition and verification," Oct. 2020, *arXiv:2010.12653*. Accessed: Aug. 03, 2021.

[17] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," Aug. 2019, *arXiv:1808.00158*. Accessed: Jan. 12, 2023.

[18] W. Xie, A. Nagrani, J. Son Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," May 2019, *arXiv:1902.10107*. Accessed: Jan. 12, 2023.

[19] J. A. C. Nunes, D. Macedo, and C. Zanchettin, "AM-MobileNet1D: A portable model for speaker recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9207519.

[20] R. Tong, L. Wang, and B. Ma, "Transfer learning for children's speech recognition," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Dec. 2017, pp. 36–39, doi: 10.1109/IALP.2017.8300540.

[21] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "Spectral modification for recognition of children's speech undermismatched conditions," in *Proc. 23rd Nordic Conf. Comput. Linguistics (NoDaLiDa)*, Reykjavik, Iceland: Linköping Univ. Electronic Press, May 2021, pp. 94–100. Accessed: Jan. 12, 2023. [Online]. Available: https://aclanthology.org/2021.nodalida-main.10

[22] E. Booth, J. Carns, C. Kennington, and N. Rafla, "Evaluating and improving child-directed automatic speech recognition," in *Proc. 12th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, May 2020, pp. 6340–6345. Accessed: Jan. 12, 2023. [Online]. Available: https://aclanthology.org/2020.lrec-1.778

[23] *End-to-End Neural Systems for Automatic Children Speech Recognition: An Empirical Study | Elsevier Enhanced Reader*. Accessed: Jan. 12, 2023. [Online]. Available: https://reader.elsevier.com/reader/sd/pii/S0885230821000905?token=B1E7CB46937771A433675363A2F64DB62FDBC8EF67723CC34A2D649560F968357D9F186DF9467BF58F35AAECAB1CDB03&originRegion=eu-west-1&originCreation=20230112172622

[24] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis," *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.
[25] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jančovič, amd M. Carey, "Speaker recognition for children's speech," in *Proc. Interspeech*, 2012, pp. 1836–1839, doi: 10.21437/Interspeech.2012-401.
[26] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, "Children's speaker verification in low and zero resource conditions," *Digit. Signal Process.*, vol. 116, Sep. 2021, Art. no. 103115, doi: 10.1016/j.dsp.2021.103115.
[27] M. Russell, "The pf-star british english childrens speech corpus," *Speech Ark Limited*, 2006.
[28] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids^2 speech corpus and recognizers," in *Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP)*, vol. 4, Oct. 2000, pp. 258–261, doi: 10.21437/ICSLP.2000-800.
[29] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids corpus," *Linguistic Data Consortium*. Philadelphia, PA, USA: Linguistic Data Consortium, 1997.
[30] S. Pradhan, R. Cole, and W. Ward, "MyST children's conversational speech," *Linguistic Data Consortium*. Philadelphia, PA, USA: Linguistic Data Consortium, 2021.
[31] *The LJ Speech Dataset*. Accessed: Jan. 12, 2023. [Online]. Available: https://keithito.com/LJ-Speech-Dataset
[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.
[33] C. Veaux, J. Yamagishi, and K. MacDonald, "SUPERSEDED—CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Rainbow Passage Which Speakers Read Out Found Int. Dialects English Arch., Apr. 2017. [Online]. Available: http://web.ku.edu/~idea/readings/rainbow.htm), doi: 10.7488/ds/1994.
[34] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, Aug. 2017, pp. 2616–2620, doi: 10.21437/Interspeech.2017-950.
[35] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, Sep. 2018, pp. 1086–1090, doi: 10.21437/Interspeech.2018-1929.
[36] P. K. O'Neill, V. Lavrukhin, S. Majumdar, V. Noroozi, Y. Zhang, O. Kuchaiev, J. Balam, Y. Dovzhenko, K. Freyberg, M. D. Shulman, B. Ginsburg, S. Watanabe, and G. Kucsko, "SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," Apr. 2021, *arXiv:2104.02014*. Accessed: Jan. 12, 2023.
[37] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," 2019, *arXiv:1904.02882*. Accessed: May 5, 2023.
[38] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," Rainbow Passage Which Speakers Read Out Can be Found Int. Dialects English Arch., Nov. 2019. [Online]. Available: http://web.ku.edu/~idea/readings/rainbow.htm, doi: 10.7488/ds/2645.
[39] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. IEEE 9th Workshop Multimedia Signal Process.*, Oct. 2007, pp. 22–25, doi: 10.1109/MMSP.2007.4412809.
[40] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 1997, pp. 473–476, doi: 10.21437/Eurospeech.1997-161.
[41] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999.
[42] R. Mugitani and S. Hiroya, "Development of vocal tract and acoustic features in children," *Acoust. Sci. Technol.*, vol. 33, no. 4, pp. 215–220, 2012, doi: 10.1250/ast.33.215.
[43] B. Ahmed, K. J. Ballard, D. Burnham, T. Sirojan, H. Mehmood, D. Estival, E. Baker, F. Cox, J. Arciuli, T. Benders, K. Demuth, B. Kelly, C. Diskin-Holdaway, M. Shahin, V. Sethu, J. Epps, C. B. Lee, and E. Ambikairajah, "AusKidTalk: an auditory-visual corpus of 3-to 12-year-old Australian children's speech," in *Proc. Interspeech*, Aug. 2021, pp. 3680–3684, doi: 10.21437/Interspeech.2021-2000.
[44] S. Shahnawazuddin, N. Adiga, H. K. Kathania, and B. T. Sai, "Creating speaker independent ASR system through prosody modification based data augmentation," *Pattern Recognit. Lett.*, vol. 131, pp. 213–218, Mar. 2020, doi: 10.1016/j.patrec.2019.12.019.
[45] V. Bhardwaj, V. Kukreja, and A. Singh, "Usage of prosody modification and acoustic adaptation for robust automatic speech recognition (ASR) system," *Revue d'Intell. Artificielle*, vol. 35, no. 3, pp. 235–242, Jun. 2021, doi: 10.18280/ria.350307.
[46] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition—The 'Ethiopian' system for the SLT 2021 children speech recognition challenge," Nov. 2020, *arXiv:2011.04547*. Accessed: Jun. 15, 2023.
[47] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral modification based data augmentation for improving end-to-end ASR for children's speech," Mar. 2022, *arXiv:2203.06600*. Accessed: Jun. 12, 2023.
[48] J. J. Burred, E. Ponsot, L. Goupil, M. Liuni, and J.-J. Aucouturier, "CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition," *PLoS ONE*, vol. 14, no. 4, Apr. 2019, Art. no. e0205943, doi: 10.1371/journal.pone.0205943.
[49] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," Nov. 2017, *arXiv:1710.10467*. Accessed: Jan. 16, 2023.
[50] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011, doi: 10.1109/TASL.2010.2064307.
[51] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4052–4056, doi: 10.1109/ICASSP.2014.6854363.
[52] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.
[53] Zchelly. (Jun. 15, 2020). *Usage of Speaker Embeddings for More Inclusive Speech-to-Text—H2020 COMPRISE*. Accessed: May 5, 2023. [Online]. Available: https://www.compriseh2020.eu/usage-of-speaker-embeddings-for-more-inclusive-speech-recognition/
[54] Resemble AI. (Jan. 15, 2023). *Resemble-AI/Resemblyzer*. Accessed: Jan. 16, 2023. [Online]. Available: https://github.com/resemble-ai/Resemblyzer
[55] *P.800: Methods for Subjective Determination of Transmission Quality*. Accessed: Jan. 17, 2023. [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I
[56] R. Jain, A. Barcovschi, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A WAV2VEC2-based experimental study on self-supervised learning methods to improve child speech Recognition.," *IEEE Access*, vol. 11, pp. 46938–46948, 2023, doi: 10.1109/ACCESS.2023.3275106.

MARIAM YAHAYAH YIWERE received the Bachelor of Science degree from the Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, in 2012, and the Master of Engineering and Ph.D. degrees from the Department of Computer Engineering, Hanbat National University, South Korea, in August 2015 and February 2020, respectively. Since October 2020, she has been working on the DTIF/DAVID project as a Postdoctoral Researcher with the College of Science and Engineering, University of Galway, Ireland. Her research interests include text-to-speech synthesis, speaker recognition and verification, sound source localization, deep learning, and computer vision.

ANDREI BARCOVSCHI received the B.Eng. degree in electronic and computer engineering from the University of Galway, in 2020, and the M.Sc. degree in artificial intelligence from the National University of Ireland Galway (NUIG), in 2021. He is currently pursuing the Ph.D. degree in artificial intelligence with the University of Galway, researching speech synthesis and conversion technologies, text-to-speech, and speech-to-text. His research interests include machine learning and artificial intelligence topics.

RISHABH JAIN (Graduate Student Member, IEEE) received the B.Tech. degree in computer science and engineering from the Vellore Institute of Technology (VIT), in 2019, and the M.S. degree in data analytics from the University of Galway, Ireland, in 2020, where he is currently pursuing the Ph.D. degree. He is also a Research Assistant with the University of Galway under the Data-center Audio/Visual Intelligence on-Device (DAVID) Project. His research interests include machine learning and artificial intelligence specifically in the domain of speech understanding, text-to-speech, speaker recognition, and automatic speech recognition.

HORIA CUCU (Member, IEEE) received the B.S. and M.S. degrees in applied electronics and the Ph.D. degree in electronics and telecom from the University Politehnica of Bucharest (UPB), Romania, in 2008 and 2011, respectively.

From 2010 to 2017, he was a Teaching Assistant and then a Lecturer with UPB, where he is currently an Associate Professor. In this position, he has authored over 75 scientific papers in international conferences and journals, served as the project director for seven research projects, and contributed as a researcher to ten other research grants. He holds two patents. In addition, he founded and leads Zevo Technology, a speech start-up dedicated to integrating state-of-the-art speech technologies in various commercial applications. His research interests include machine/deep learning and artificial intelligence, with a special focus on automatic speech and speaker recognition, text-to-speech synthesis, and speech emotion recognition.

Dr. Cucu was awarded the Romanian Academy Prize "Mihail Drăgănescu," in 2016, for outstanding research contributions in Spoken Language Technology, after developing the first large-vocabulary automatic speech recognition system for the Romanian language.

PETER CORCORAN (Fellow, IEEE) is currently the Personal Chair of Electronic Engineering with the College of Science and Engineering, University of Galway, Ireland. He was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He has more than 600 cited technical publications and patents, more than 120 peer-reviewed journal articles, and 160 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction, and facial detection. He is also a member of the IEEE Consumer Technology Society for more than 25 years and the Founding Editor of *IEEE Consumer Electronics Magazine*.

Appendix H

Data Center Audio/Video Intelligence on Device (DAVID) - An Edge-AI Platform for Smart-Toys.

Authors: Gabriel Costache (GC), Francisco Salgado (FS), Cosmin Rotariu (CR), George Sterpu (GS), Rishabh Jain (RJ), and Peter Corcoran (PC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	GC: 60%, FS: 10%, CR: 10%, GS: 10%, RJ: 5%, PC: 5%
Experiments and Implementation	GC: 30%, FS: 20%, CR: 20%, GS: 20%, RJ: 10%
Background	GC: 50%, FS: 10%, CR: 10%, GS: 10%, RJ: 10%, PC: 10%
Manuscript Preparation	RJ: 60%, PC: 20%, GC: 5%, FS: 5%, CR: 5%, GS: 5%

Data Center Audio/Video Intelligence on Device (DAVID) - An Edge-AI Platform for Smart-Toys

Gabriel Cosache
DTS (office of CTO),
Xperi Corporation
Galway, Ireland
gabriel.costache@xperi.com

Francisco Salgado
DTS (office of CTO),
Xperi Corporation
Galway, Ireland
francisco.salgado@xperi.com

Rishabh Jain
C3 Imaging Research Center
University of Galway
Galway, Ireland
rishabh.jain@universityofgalway.ie

Cosmin Rotariu
DTS (office of CTO),
Xperi Corporation
Galway, Ireland
cosmin.rotariu@xperi.com

George Sterpu
DTS (office of CTO),
Xperi Corporation
Galway, Ireland
george.sterpu@xperi.com

Peter Corcoran
C3 Imaging Research Center
University of Galway
Galway, Ireland
peter.corcoran@universityofgalway.ie

2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) | 979-8-3503-2797-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/SpeD59241.2023.10314915

***Abstract*—An overview is given of the DAVID Smart-Toy platform, one of the first Edge-AI platform designs to incorporate advanced low-power data processing by neural inference models co-located with the relevant image or audio sensors. There is also on-board capability for in-device text-to-speech generation. Two alternative embodiments are presented – a smart Teddy-bear and a roving dog-like robot. The platform offers a speech-driven user interface and can observe and interpret user actions and facial expressions via its computer vision sensor node. A particular benefit of this design is that no personally identifiable information passes beyond the neural inference nodes thus providing inbuild compliance with data protection regulations.**

Keywords—smart-toy, speech user interface, neural networks, Edge-AI, privacy-by-design, human-computer interface

I. Introduction

Edge-AI is an emerging concept implying a migration of computational intelligence and associated data processing from cloud repositories to occur closer to the source of data. In some interpretations, it implies data processing on a local smartphone or hub device, but the ultimate goal of Edge-AI is on-device processing. Typically, the artificial intelligence element is a neural model that leverages recent advances in processing images, speech, or other raw sensor data sources. However, as the capabilities of most embedded inference chipsets are relatively limited and still require significant compute power [1], [2] most designs implement quite limited or specific functionality [3]. The inference requirements of advanced computer vision and automated speech models further limit the capabilities of Edge-AI implementations [2], [4]–[6]. Fortunately a new generation of specialized neural accelerators [7]–[9] are capable of running larger neural models and even combinations of models, as we shall see.

Data privacy has also become a significant consideration, especially for consumer devices and services [10]–[13]. More specifically, smart-toys have led to much controversy when they collect personally identifiable data from children [14], [15]. Clearly, privacy is of particular importance when dealing with children. Thus, data security has been a primary concern for the platform and a key design requirement was to eliminate sending any data that might be considered personally identifiable beyond the sensor nodes. Due to the scope of the *General Data Protection Regulations* (GDPR) in the European field [16]. This implies that all image or speech data should be processed on the sensor boards.

II. System Design and Architecture Overview

The DAVID project required a large-scale development effort over three years, so it is not feasible to capture the many details of the design, test, and final implementation of the hardware designs. Here we present the final working system, focusing on the key privacy-by-design aspects.

A. The ERGO Neural Accelerator

The DAVID platform design was inspired by the recent availability of low-power inference chipsets such as ERGO [17]. It provides ultra-low-power capabilities while delivering significant computational capabilities – of the order of 50 TOPS/Watt. Thus, for the smart-toy use case, a computational loading of 2-3 TOPS is feasible for a power budget of 50 mW – approximately the same power consumption as two light-emitting diodes.

B. System Architecture Overview

The system hardware layout is provided in *Figure 1*. This comprises three dedicated inference nodes, one with an onboard camera enabling a computer vision node; the second with a microphone to provide an audio sensing node – the primary function will be operating the speech interface with the user, and the third board is linked to a speaker to enable the smart toy to generate neural voice output. This third board is not a sensor board but is needed to close the loop on a speech-based user interface. This demonstrates the capability of Edge-AI to also generate data outputs.

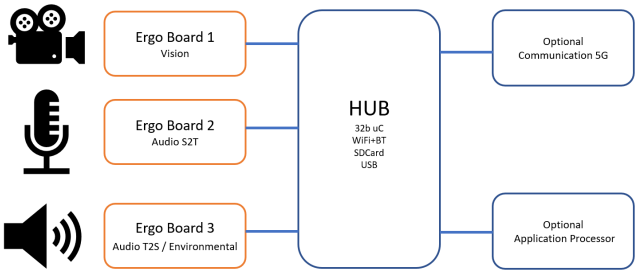


Figure 1: *System Hardware Architecture for DAVID*

These three inference boards are connected via an I2S bus to a central, low-power microcontroller (MCU) hub. This is also an ultra-low-power design but consumes more power than the sensor nodes unless placed in deep shutdown mode.

C. The Hub Board

The Hub board, shown in *Figure 2*, was designed to leverage the many state-of-the-art features available on today's smartphones including IMU subsystem, programmable wireless connectivity, and JTAG. It features an STM32H7 low-power MCU which can support full 32-bit OS. This was selected for his flexibility and support for all peripherals we have considered combined with its low power requirements. The Hub board's main purpose is to connect the Ergo boards and manage communication. The Hub should only be live when the Ergo boards are sensing events that require the Hub board to provide input to any of the peripherals attached to it. It will be in sleep mode most of the time. This helps to ensure very low power requirements for the overall system platform.

TABLE 1: CONNECTIVITY, MEMORY & COMPUTATIONAL CAPABILITIES OF THE DAVID PLATFORM (HUB + NODES)

CONNECTIVITY	HUB	NODE	COMPUTATION	HUB	NODE
I2S (Tx, Rx), I2C	X	X	Ergo, 55TOPs/Watt x3		X
MIPI and Parallel		X	Arc CPU\DSP)		X
SPI & QSPI	X	X	STM32 (Arm M7)	X	
GPIO (32 bit)	X	X	ESP32 (Xtensa LX6)	X	
FTDI (JTAG, UART)	X				
WiFi/BT	X				
USB OTG	X				
Memory					
16MB QSPI Flash		X			
128MB QSPI Flash	X				
32MB SRAM	X				
448 KB ROM	ESP32				
520 KB SRAM	ESP32				
SDCard	X				

Figure 2: Top and Bottom views of the DAVID Hub board.

D. The Sensor Nodes

The sensor nodes are designed to leverage the capabilities of the ERGO chipset. They feature an onboard MIPI bus to allow for fast real-time data transfer from an onboard camera or other high-bandwidth sensors and to allow fast data transfers between ERGO and the onboard memory subsystem. Each node also has built-in IMU and I/O ports to support a range of different sensor peripherals. For the initial DAVID proof-of-concept (PoC) three different sensor nodes were configured:

Vision Node: This node features a wide-field QVGA MIPI camera. This choice was made as a trade-off between the resolution required to implement most of the computer vision algorithms and the complexity of the neural architectures. Most of the selected neural algorithms operate well with QVGA input, and this allows more neural models to be incorporated and operate in parallel. The MIPI architecture is scalable to a higher (or lower) resolution camera sensor as required.

Audio Node: This node features two microphone inputs – one is a low-quality consumer-grade microphone, but a higher quality stereo microphone is also incorporated and used for initial demos and to gather test input data with different quality.

Speaker Node: This node was configured with several different digital output speakers for testing. From our experience, it is difficult to achieve significant loudness in the output without added amplification, but this has consequences for power consumption. To date, we did not find a good low-power solution and the PoC speech output can be difficult to hear in a noisy environment. However, ongoing improvements in sound output components from smartphones should help solve this issue at a sensible price/performance point in the near future.

E. The DAVID Platform

The hub board and up to three Inference nodes are designed to be assembled into a single platform unit that can be designed into a proof-of-concept smart toy. This is illustrated as a block diagram in *Figure 3, below.* Various on-board connectors simplify connecting the electronic platform to externally mounted cameras, microphones, sensors, speakers, or other equivalent peripherals that consume or generate data. The design is intended to be as generic as possible to facilitate incorporation into demonstrators for different consumer devices/products. A picture of the two sides of the hub board is shown in *Figure 2,* opposite.

Naturally, this platform is not intended for mass-market manufacturing, only for proof-of-concept (PoC) designs. Ultimately the different system components would be incorporated into a single chip for a particular product design. Here the selection of ARM-based MCU and mass-market camera, microphone, and speaker peripherals will simplify the transition from PoC to the final mass-market product. The fully assembled system is shown in *Figure 4,* below.

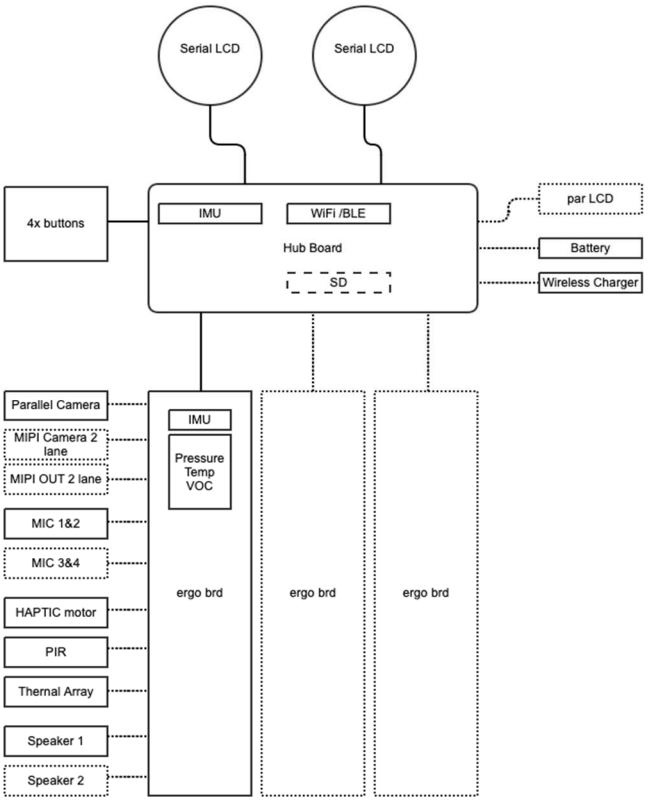


Figure 3: Detailed System Block Diagram.

Figure 4: *DAVID Platform – Hub with 3 Inference node daughter boards fully assembled for placement in a smart-toy.*

III. Edge-AI Models and the System User Interface

In this section, we take a look at the neural models implemented on each of the node boards and discuss some of the challenges in training and preparing neural models to run on the DAVID platform.

A. Porting of NN Models to the Edge-AI Platform

Prior to the compilation and compression of a neural model onto the ERGO is it important to simplify the original neural model (or composite model if several models are to combine on a single ERGO board). This is a complex process and typically involves a mix of layer quantization and node pruning. In some cases – a well-known example is the YOLO object detection framework – there may be a range of model sizes (tiny, small, regular, and large models) with varying levels of performance/complexity to provide a repeatable starting point. Here, due to space limitations, we only comment that this process is empirical and there is a need to balance the desired levels of performance with the capabilities of the ERGO tools to covert models for loading on ERGO. In theory, any model that can run on *PyTorch* can be converted, but in some cases, it was found that models simply could not achieve acceptable levels of performance.

B. The Vision Node

This sensing node is perhaps the most advanced as it leverages several decades of computational imaging experience in the CTO office of Xperi [18], [19]. This node incorporates a multi-function neural model that can perform a range of vision tasks. An overview of one representative version is shown in *Figure 5*, but this inference module can be configured in various alternative arrangements, depending on the specific application. Here the neural model can detect both facial, hand, and body regions and a set of analytics is performed on each of these. In addition, body landmarks, hand gestures (from a fixed set of classes), and facial characteristics are output. The facial analysis is the most sophisticated, including orientation, a landmark mesh, facial expression (from a fixed set of classes), and facial embedding data that can be used to authenticate the user. All of these data outputs are available in numeric form to the hub board but no sensitive biometric data is exported beyond the vision node.

In addition, this example also includes a neural video encoder that allows an encoded video stream to be set over a secure wireless link (Bluetooth) to a parent's phone. The video encoding can only be viewed on a paired app running on the phone and requires a custom decoder to view the video. This is the only export of PID data outside of the platform and is provided to show that secure parental access can be provided. The functionality illustrated in *Figure 5* is available in real-time at frame rates of 30 fps. The total power cost is 100 mW. By deactivating the secure streaming functionality this can be reduced significantly. It is worth noting that the system architecture allows for fast reprogramming of each Inference node. Typically each can be re-flashed with a separate functionality in a few 10's of milliseconds. Thus this node could be re-flashed with a different set of functionality to support a specific play activity or to switch between single-player and multi-player use cases. Multiple Inference models can be stored in flash, or uploaded via a secure app to the system allowing for additional flexibility in the uses of the platform.

C. The Text-to-Speech (TTS) Node

The TTS Edge AI pipeline is composed of two modules: a spectrogram network and a vocoder. The spectrogram network was initially adapted from the well-known Tacotron model [20]. However, this architecture required off-node computations in the hub MCU and the model was slow to re-converge. Due to these challenges later versions of the TTS model switched to explore the FastSpeech end-to-end model [21], [22] and later several optimized versions of this model [23]. For the Vocoder, several alternatives were explored with the chosen architecture being HiFiGAN [24]. An overview of TTS architectures can be found here [20].

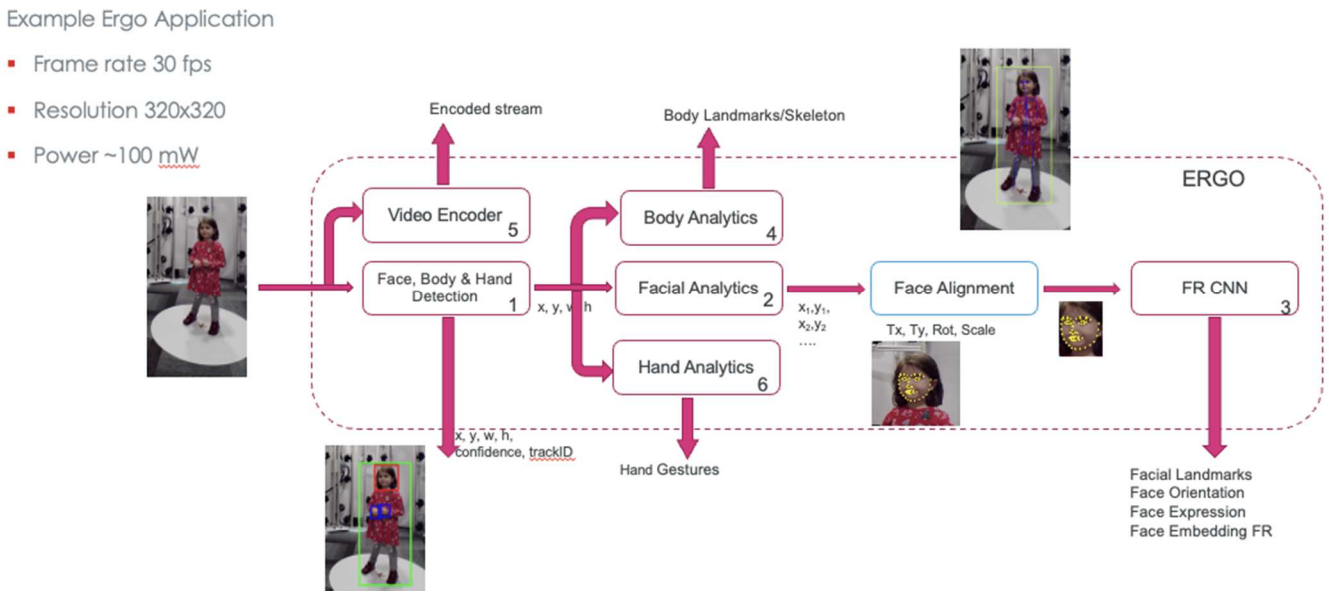


Figure 5: *The multi-CNN model structure implemented in the Computer Vision Sensor Node of DAVID.*

The optimized spectrogram network, shown in *Figure 6*, is a fully convolutional non-autoregressive network. It computes an intermediate log mel-scale spectrogram representation of the whole input text in parallel. The network takes the text characters as inputs, avoiding a separate phoneme conversion module. The network models the durations of the encoded characters directly prior to upsampling them to match the spectrogram length. During training, it uses the self-alignment module from FastPitch [21], which removes the need for external precomputed alignments between the text and the corresponding audio and easily scales to new datasets.

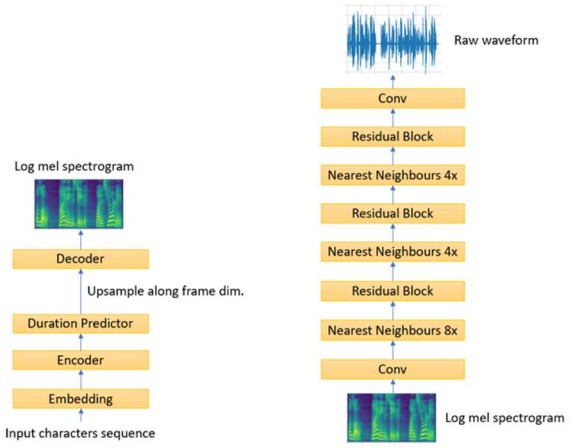


Figure 6: *Left-side: The Spectrogram network diagram. Right-side: The Vocoder generator diagram.*

The vocoder follows the generator architecture of [24] interleaving fully convolutional residual blocks with nearest neighbor's upsampling layers. The convolutional kernel sizes and the receptive field of the network have been tuned for inference on the hardware at the chosen output sampling rate. During inference, the vocoder uses a sliding window approach to reduce activation memory. The network takes as input non-overlapping parts of the log mel-scale spectrogram and computes the corresponding audio waveforms, which are then concatenated to produce the complete speech waveform.

Both modules are trained separately. The spectrogram network is trained by optimizing the mean-squared error between the log mel-scale spectrogram of the predicted and ground-truth signals as well as the alignment error. The vocoder is trained using the loss proposed in [24].

D. The Automatic Speech Recognition Node

SpeechNet is a fully convolutional neural architecture designed for real-time speech recognition on the ERGO hardware. In contrast with related convolutional models in ASR, such as Jasper [22] or QuartzNet [23] SpeechNet relies on a considerably shorter audio context and uses small kernel sizes. The shorter receptive field is achieved by reducing the network depth at the expense of the layer width, where multiple convolutions are executed in parallel, similar to the building block of the Inception architecture [24].

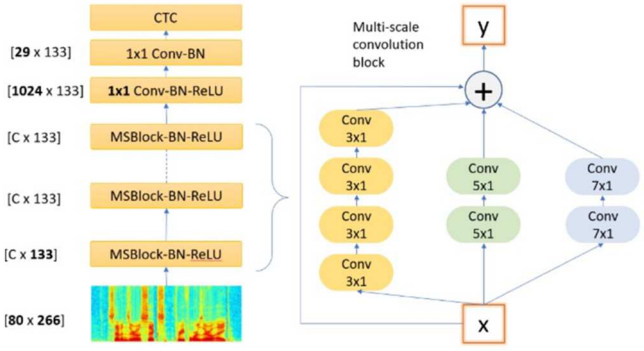


Figure 7: *SpeechNet1 diagram. Left: Overall structure. Right: Zoom on a Multi-scale convolution block (MSBlock)*

The basic building block of SpeechNet is shown on the right-hand side of the diagram in *Figure 7* and consists of several Convolutional layers, each followed by Batch Normalization (BN) and Rectified Linear Unit (ReLU) non-linearities. For brevity, the BN and ReLU operations are not included in the diagram. The block input x is transformed by three different convolutional paths made of two or more Conv-BN-ReLU blocks, each path using a different kernel size. Furthermore, with the exception of the first MSBlock in

the network, all MSBlocks add a residual connection from input to output. The input feature to SpeechNet is a normalized time-frequency log mel-scale spectrogram representation of an audio signal. The model outputs an array of 29 posterior probabilities defined over an inventory of 26 letters in the English alphabet, blank space, apostrophe, and the CTC skip/BLANK token, and is trained using the Connectionist Temporal Classification (CTC) objective function [25] within a supervised learning framework.

For simplicity, SpeechNet uses characters as modeling units, lessening the burden of generating richer transcriptions. Through an appropriate parametrization of the network about depth, kernel sizes, strides, and padding, SpeechNet1 achieves a constant algorithmic latency of 1.3 seconds, relying on 3.325 seconds of audio context. Building upon SpeechNet, we developed a more generic SpeechNet2 network replacing the log mel-scale spectrogram feature pre-processor with a fully learnable convolution-based front-end. This front-end is designed to accept audio waveforms as inputs, which are apriori mu-law quantized on 8 bits to satisfy the input data type requirements of the ERGO processor.

IV. Smart Toy Proof-of-Concept

To demonstrate practical use cases for the DAVID platform we have developed two different example use cases. The original use case was a soft toy such as a Teddy bear that a child can talk with and engage in different play activities. A second example of a mobile robot was developed to take advantage of toy mobility to enhance some of the play activities, demonstrate additional functionality, and explore if active mobility can help improve user engagement with the smart toy. Here we present some technical details on how the underlying Edge-AI platform was integrated into each PoC.

A. The DAVID Smart Teddy-Bear

The initial proof-of-concept (PoC) embodiment for the DAVID smart toy is a Teddy-Bear, or more correctly a Panda. This was chosen as many children have a special cuddly toy that they develop a close attachment to. Having a toy with a full speech interface that children can talk to (see Figure 9) provides an interesting toy variant to test and evaluate.

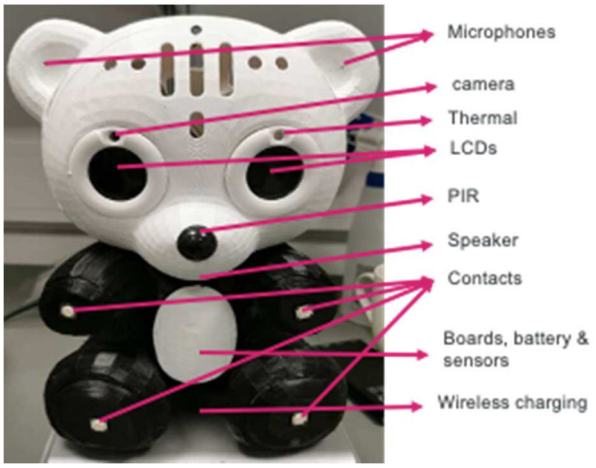


Figure 8: *The DAVID Teddy Bear Design*

From a technical perspective, it is also less challenging to design and implement a static toy platform. For this proof-of-concept, the focus is on providing a more sophisticated speech interface with interactive activities such as games and storytelling. The bear casing is designed to be 3D printable. A special 3D printing material is added later to provide a synthetic soft fur over the hard casing of the final toy. The toy platform hosts various sensing and UI elements as shown in *Figure 8*. It can be seated on a wireless charging cradle, but in normal operation, it can run for more than 1 week on the internal, rechargeable battery pack.

Figure 9: *Testing Smart-Toy interactivity in the lab - the toy's eyes follow the user and can provide indications of emotional states.*

B. The DAVID Mobile Rover

A second design PoC embodiment is provided by a mobile rover. This is a pet-like platform that can move around and execute a variety of different movement patterns, including 180 and 360-degree rotations. It also adds a head-like data screen that can nod and antennae that can perform various movements. These can be used to express happiness or indicate the toy is upset or confused.

Several variants of this mobile design have been tested and improved. This platform can employ a less sophisticated speech interface for many demonstration tasks and games. Simple wake words and simple speech commands can enable interesting play experiences. As this variant of the toy mimics a robo-pet the gameplay experiences can be simpler. Thus, chasing a ball, or dancing while playing some music can provide a very entertaining experience. Emotional cues can be simpler, again mimicking a pet puppy. Many of the demo activities for the mobile DAVID embodiment were noted by the project team as being more fun to work on than the sophisticated interactions expected in activities for the cuddly toy.

V. Conclusions

The DAVID smart-toy platform provides an interesting overview of how future Edge-AI platforms are likely to evolve. The migration of much of the intelligence onto the sensing nodes can allow designers and developers to focus on the gaming or activity logic without worrying about how to integrate computer vision algorithms or speech recognition elements. Speech is analyzed and generated from the underlying text representations and computer vision algorithms can provide authentication and analysis data in simplified forms.

Perhaps more important here are the data privacy benefits. As both speech and face data are regarded as biometrics and thus classified as personally identifiable data (PID) developers would be exposed to GDPR compliance issues. By embedding

the processing of this sensor data onto the ERGO node boards this removes the need for GDPR compliance for smart-toy designers.

ACKNOWLEDGMENTS

This work was supported by the DAVID project of the Disruptive Technologies Innovation Fund (managed by the Department of Enterprise, Trade and Employment and administered by Enterprise Ireland) and the College of Science & Engineering Ph.D. Research Scholarship at the University of Galway.

REFERENCES

[1] M. A. Farooq, W. Shariff, and P. Corcoran, "Evaluation of Thermal Imaging on Embedded GPU Platforms for Application in Vehicular Assistance Systems," *IEEE Trans. Intell. Veh.*, pp. 1–1, 2022, doi: 10.1109/TIV.2022.3158094.

[2] D. Bigioi and P. Corcoran, "Challenges for edge-ai implementations of text-to-speech synthesis," in *2021 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, 2021, pp. 1–6.

[3] W. Yao, V. Varkarakis, G. Costache, J. Lemley, and P. Corcoran, "Towards Robust Facial Authentication for Low-Power Edge-AI Consumer Devices," *IEEE Access*, pp. 1–1, 2022, doi: 10.1109/ACCESS.2022.3224437.

[4] B. Sudharsan, "On-Device Learning, Optimization, Efficient Deployment and Execution of Machine Learning Algorithms on Resource-Constrained IoT Hardware," 2022.

[5] B. Sudharsan, S. Malik, P. Corcoran, P. Patel, J. G. Breslin, and M. I. Ali, "OWSNet: Towards Real-time Offensive Words Spotting Network for Consumer IoT Devices," in *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*, Jun. 2021, pp. 83–88. doi: 10.1109/WF-IoT51360.2021.9595421.

[6] B. Sudharsan, P. Patel, A. Wahid, M. Yahya, J. G. Breslin, and M. I. Ali, "Porting and execution of anomalies detection models on embedded systems in iot: Demo abstract," in *Proceedings of the international conference on internet-of-things design and implementation*, 2021, pp. 265–266.

[7] P. Pandey *et al.*, "Challenges and opportunities in near-threshold dnn accelerators around timing errors," *J. Low Power Electron. Appl.*, vol. 10, no. 4, p. 33, 2020.

[8] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "AI and ML accelerator survey and trends," in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, IEEE, 2022, pp. 1–10.

[9] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey and benchmarking of machine learning accelerators," in *2019 IEEE high performance extreme computing conference (HPEC)*, IEEE, 2019, pp. 1–9.

[10] P. M. Corcoran, "A privacy framework for the Internet of Things," in *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, Dec. 2016, pp. 13–18. doi: 10.1109/WF-IoT.2016.7845505.

[11] P. Corcoran, "Privacy challenges for smart-cities: The challenge of IoT camera uberveilance," IEEE World Forum on Internet of Things, 2019.

[12] J. Hinds, E. J. Williams, and A. N. Joinson, "'It wouldn't happen to me': Privacy concerns and perspectives following the Cambridge Analytica scandal," *Int. J. Hum.-Comput. Stud.*, vol. 143, p. 102498, 2020.

[13] A. Karale, "The challenges of IoT addressing security, ethics, privacy, and laws," *Internet Things*, vol. 15, p. 100420, 2021.

[14] E. Taylor and K. Michael, "Smart toys that are the stuff of nightmares," *IEEE Technol. Soc. Mag.*, vol. 35, no. 1, pp. 8–10, 2016.

[15] J. Valente and A. A. Cardenas, "Security & privacy in smart toys," in *Proceedings of the 2017 Workshop on Internet of Things Security and Privacy*, 2017, pp. 19–24.

[16] H. Li, L. Yu, and W. He, "The impact of GDPR on global technology development," *Journal of Global Information Technology Management*, vol. 22, no. 1. Taylor & Francis, pp. 1–6, 2019.

[17] "Perceive – Transform Sensing into Perceiving." https://perceive.io.

[18] P. Corcoran, P. Bigioi, E. Steinberg, and A. Pososin, "Automated in-camera detection of flash-eye defects," *IEEE Trans. Consum. Electron.*, vol. 51, no. 1, pp. 11–17, 2005.

[19] P. M Corcoran, P. Bigioi, and F. Nanu, "Advances in the detection & repair of flash-eye defects in digital images-a review of recent patents," *Recent Pat. Electr. Electron. Eng. Former. Recent Pat. Electr. Eng.*, vol. 5, no. 1, pp. 30–54, 2012.

[20] N. Kaur and P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: a review," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 5837–5880, Jul. 2023, doi: 10.1007/s10462-022-10315-0.

[21] R. Badlani, A. \Lańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One TTS alignment to rule them all," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6092–6096.

[22] J. Li *et al.*, "Jasper: An end-to-end convolutional neural acoustic model," *ArXiv Prepr. ArXiv190403288*, 2019.

[23] S. Kriman *et al.*, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6124–6128.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

Appendix I

Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing.

Authors: Dan Bigioi (DB), Hugh Jordon (HJ), Rishabh Jain (RJ), Rachel McDonnell (RM) and Peter Corcoran (PC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	DB: 80%, RM: 10%, PC: 10%
Experiments and Implementation	DB: 80%, HJ: 10%, RJ: 10%
Background	DB: 60%, HJ: 10%, RJ:10%, RM:10%, PC:10%
Manuscript Preparation	DB: 70%, HJ: 10%, RJ: 10%, RM: 5%, PC: 5%

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 December 2022, accepted 14 December 2022, date of publication 20 December 2022,
date of current version 28 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3231137

RESEARCH ARTICLE

Pose-Aware Speech Driven Facial Landmark Animation Pipeline for Automated Dubbing

DAN BIGIOI[1], (Graduate Student Member, IEEE), HUGH JORDAN[2], RISHABH JAIN[1], (Member, IEEE), RACHEL MCDONNELL[2], AND PETER CORCORAN[1], (Fellow, IEEE)

[1]School of Electrical and Electronics Engineering, National University of Ireland, University of Galway, Galway, H91 TK33 Ireland
[2]Trinity College Dublin, University of Dublin, Dublin 2, D02 PN40 Ireland

Corresponding author: Dan Bigioi (d.bigioi1@nuigalway.ie)

This work was supported by the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant 18/CRT/6224.

ABSTRACT A novel neural pipeline allowing one to generate pose aware 3D animated facial landmarks synchronised to a target speech signal is proposed for the task of automatic dubbing. The goal is to automatically synchronize a target actors' lips and facial motion to an unseen speech sequence, while maintaining the quality of the original performance. Given a 3D facial key point sequence extracted from any reference video, and a target audio clip, the neural pipeline learns how to generate head pose aware, identity aware landmarks and outputs accurate 3D lip motion directly at the inference stage. These generated landmarks can be used to render a photo-realistic video via an additional image to image conversion stage. In this paper, a novel data augmentation technique is introduced that increases the size of the training dataset from N audio/visual pairs up to NxN unique pairs for the task of automatic dubbing. The trained inference pipeline employs a LSTM-based network that takes Mel-coefficients as input from an unseen speech sequence, combined with head pose, and identity parameters extracted from a reference video to generate a new set of pose aware 3D landmarks that are synchronized with the unseen speech.

INDEX TERMS Machine learning, computer vision, lip synchronization, talking head generation, automatic dubbing, audio driven deep fakes, artificial intelligence.

I. INTRODUCTION

Automatic speech dubbing is an area of great interest to the entertainment sector as not only is it relevant to the task of automatic dubbing for movies, television, and videos in general, it is also applicable to speech-based animation pipelines for video game characters, CG animated movies, and increasingly, personal avatars within the realm of virtual reality.

Automatic audio-visual speech dubbing is a topic which falls under the broader field of talking head generation, or talking heads for short. A talking head video is a video which contains one subject talking directly to the camera. The goal of talking head generation is either to generate a photo-realistic talking head video from a static reference image and target audio source (image-based methods), or in the case of this paper and the task of automatic dubbing / speech driven video editing, to modify an existing video based on a new target audio clip (video-based methods).

The meteoric rise in popularity of deep learning over the last decade has in turn lead to a surge in interest towards talking head generation and its associated sub tasks such as dubbing, video editing, and video generation. Numerous approaches have been suggested over the last five years, each one looking to advance the state of the art within the field of talking heads. For the vast majority of image-based methods (where a video is generated from a single reference image + audio), a neural network is trained to generate the lip movements and facial expressions from audio, while a second network is trained to generate the head pose information. Likewise for most video-based methods (where the content of an already existing video is modified based off the audio), a single network is used to generate the lip movements onto a static face mesh, which then gets fitted on top of landmarks

The associate editor coordinating the review of this manuscript and approving it for publication was Ángel F. García-Fernández.

extracted from each frame of the video before rendering. For both cases, these pipelines are quite complex, and there is a need for simpler, more intuitive approaches such that artists can make better use of these technologies.

Speech dubbing itself is a highly complex task, as not only does one need to generate accurate lip and jaw motion to match the target speech signal, special care must be taken to not diminish from the actors visual performance. Factors such as the actors facial expressions, head movements, and mannerisms, must be kept as close to the original performance as possible such that the only difference between the dubbed video, and the original is the motion of the lips and jaw in response to the new target audio.

The aim of this paper is to test the feasibility of a novel 3D landmark pipeline that outputs pose, and identity aware talking head landmarks directly in one forward pass given an unseen target audio speech signal, and video to be modified. This differs from other approaches in the literature which typically generate moving lips onto an identity removing, static fixed head before aligning the lips with the desired head pose in a later step. In these approaches the static mesh then must be given identity specific information such as head pose and general head movement by either a separate network that generates artificial head pose sequences (when generating a video from a static image), or by extracting head pose information from the reference video and refitting the static mesh to match it. More commonly when modifying video, an intermediate 3D model is used to generate the desired facial animations, before rendering back to photorealistic frames like in [1]. Typically these methods and techniques are a lot more complex to implement and run than landmark-based solutions. An aim of this work therefore is to take the first steps towards a landmark based video modifying pipeline that may serve as a lighter, simpler, and more practical tool for animation. To this end, two main contributions are made as part of this work:

- A novel lightweight LSTM-Based Model capable of generating pose and identity aware 3D landmark sequences driven by a target audio speech signal and source video clip.
- A novel data augmentation technique for de-correlating lip, jaw, and head motion, making the generation of pose-aware landmarks possible directly at inference.

The rest of this paper is organized as follows. In section 2 a review of recent relevant works in the literature is provided to give context for this paper. To this end a concise taxonomy of papers and methods within the field of audio driven talking head generation is presented. In section 3 the methodology of the approach is reviewed, discussing the contribution of the paper in depth, and detailing the data processing methods, the network architecture, the training set up, and experiments. In section 4 the results are presented and discussed. In section 5 societal impact and ethical considerations of the work are discussed before the conclusion of the paper.

II. RELATED WORKS

Talking head generation is a topic which falls under the wider umbrella of ''Deep Fakes'', where the goal is to generate realistic fake content of a target person. There are many different approaches for generating ''Deep Fakes'', making a detailed literature review of the topic challenging. Here the scope of the related works section is limited to research with a focus on facial animation and motion driven directly from a speech sequence - audio driven talking heads. For a more thorough review of the literature surrounding the topic of ''Deep Fakes'' the reader is directed to [2] as it provides a comprehensive overview of the field and the main methods for generating fake content.

Following a thorough review of the literature surrounding audio-driven talking heads, several interesting pipelines were identified that could be applied to the task of automatic speech dubbing. These pipelines can be broadly classified into two over-arching approaches: Structural approaches [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30] transform the input image/video and audio into an intermediate structural representation (typically, 2D facial landmarks, or a 3D mesh) that is used as input to a neural renderer to generate a photo-realistic talking head sequence. Image reconstruction approaches [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41] leverage direct image reconstruction techniques and latent feature learning to generate a photo-realistic sequence from a target audio and reference image/video.

While there are many approaches out there that generate lip movements from audio such as [19], [22], [26], and [42], none of these approaches generate pose-aware landmarks in a single forward pass, instead generating the lip movements onto a static face shape, introducing head movement at a later step. In this paper it is argued that this is inefficient, and can be done directly at inference time through a simple data augmentation. Creating a faster pipeline, with less moving parts, that lends itself better to real time usage.

This work is inspired by and extends the methodology presented in [19] and [20], which are approaches that take in a target audio clip as input, and generate fixed (no head pose, just lip movement) 2D talking face landmarks as output. The approach presented in this paper allows one to generate 3D talking face landmarks that maintain the head pose and identity of the original speaker, while accurately driving the lips from the target audio.

This work is also comparable to [22], which is an approach used to generate talking head animations given a single target image and audio clip. Specifically, one can compare the model in this work to their landmark prediction network which disentangles the audio into content and speaker identity embeddings. These embeddings are used to predict the landmark displacements, which are then rendered into either photo-realistic or animated frames.The approach presented in this paper works on modifying an existing video rather than generating a new one from a single image, and modifies

the landmarks based on Mel Coefficients extracted from the audio sequence that are fed into the network.

The aim of talking head generation is to generate a photo-realistic audio driven talking head video in which the facial movements of the talking head are naturally synchronized with the target speech. Note that ''audio driven talking head video'' is used as a blanket term to encompass all works related to generating facial motions and animations driven by audio, regardless of whether it is modifying a preexisting video or animating a static image.

Most of the works referenced in this section, can be classified into one of two fundamental approaches for the task of audio driven talking head generation: structural based methods [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], and image reconstruction-based methods [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41].

Structural Based Methods: These are approaches where the input image, video, or audio are transformed into an intermediate structural representation of some sort such as a 3D model / mesh [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] a sequence of facial landmarks [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], and more recently a sequence of dense motion fields [18], [30]. These are then used as a training feature for an underlying network that takes these structural sequences as input to render a photo-realistic video. These methods are the most relevant to this paper, specifically landmark based ones, as this work introduces a novel way of generating pose aware 3D facial landmark sequences from a preexisting video sequence and target audio clip.

Image Reconstruction Methods: These are the approaches which use pure image reconstruction techniques and latent feature learning [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41]. One could consider these as true ''End to End'' approaches, essentially passing the target image/video and audio through a generative neural network, outputting the synchronized talking head video directly. Other Methods: These are approaches which do not strictly fall within the two classes above, that are still highly relevant to this field and worth mentioning. Approaches such as [42] and [43] which are audio driven models trained to animate face rigs through visemes. Or [44] that can generate dynamic neural radiance fields from audio and using them to synthesize photorealistic talking head videos.

It is also worth noting that each of the categories mentioned above can be further broken down into whether they are image or video-based methods.

- *Image Based Methods:* The goal is to animate a cropped facial image given an input image/limited number of frames as a reference, and an audio clip.
- *Video Based Methods:* Where the goal is to alter the lip movements and facial expressions of an already existing video so that they are synchronized with a new audio clip. Generally, the videos are full frame, containing the background, face, neck, and torso regions, not just the cropped face, unlike the image-based methods. The work presented in this paper is a video-based approach, as it seeks to modify existing video based on a new target speech audio clip.

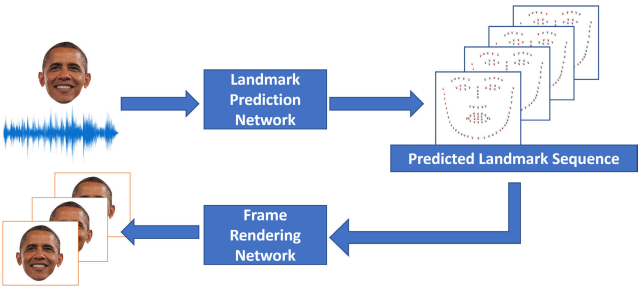


FIGURE 1. Typical landmark based pipeline.

A. LANDMARK BASED METHODS

Typically, with landmark-based methods [16], [17], [18], [19], [20], [21], [22], [23], [24], [26], [27], [28], [29], the goal across all approaches is to generate frame by frame a set of predicted facial landmarks based on a reference image/video, driven by an audio clip. The predicted landmark sequence is then passed through a separate rendering network to generate the photorealistic video frames required for the final output. Figure 1 is a simplified example of what a typical pipeline looks like, note the two main components, the landmark prediction, and the frame rendering modules. As the contribution of this paper is a novel landmark generation technique, this section focuses discussion on the various landmark prediction modules across the literature, with less emphasis on the rendering side of things.

It stands to reason that there is a lot of variation across approaches regarding the most effective method of construction for the landmark prediction module. Most modules in the literature can be grouped according to the following design choices:

- *Audio input pre-processing:* Some approaches take in phoneme labels extracted from audio like in [16]. Others extract Mel spectrograms or MFCCs from the audio first which are then fed into the predictor such as the approaches taken by [18], [19], [20], [23], [24], [25], [26], [28], and [29]. Audio embeddings obtained from trained speech to text modules such as the approaches employed [21] and [27] have also been tried, along with methods that take in custom audio embeddings such as [17] and [22]. For the approach within this paper, mel-coefficients are extracted from the audio and fed in as input features to the network. They were chosen as they are quite easy to extract compared to other audio features used by some of the approaches mentioned above, and they are immensely popular in classical speech related tasks such as text to speech, speaker recognition, and automatic speech recognition.
- *The underlying network architecture:* Some approaches such as [17], [19], [20], [21], [22], [23], [25], [26], and [28]

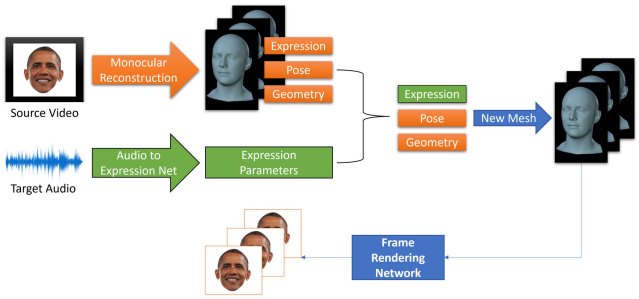


FIGURE 2. Typical 3D Model based pipeline.

employ a recurrent neural architecture and others such as [16], [18], [24], [27], and [29] use feed-forward designs. While feed-forward architectures are typically faster, a recurrent, lstm-based approach very similar to [19] and [22] was chosen for this paper. The idea was that by using a recurrent architecture, the network would learn the temporal dependence associated with audio and its output lip movements, generating a higher quality of lip movements.

• *Generating the Output Landmarks:* Some approaches in the literature generate the output landmarks using a static face mesh with moving lips that needs to then be fitted to a target video such as [16], [17], [19], [20], [23], [26], [27], and [28] while others such as [18], [21], [22], [24], [25], and [29] generate the head pose information using one network, and a second network generates the lip movements, combining both to have a pose inclusive face mesh. This paper's approach differs to these as it uses a single network trained to generate 3D pose aware landmarks synchronized to audio as described by Figure 3. This is done to simplify the overall landmark generation pipeline for faster inference speeds, and doing so allows for the generation of more accurate landmarks as less information is lost through extensive normalisation of the ground truth.

Often, the rendering modules are variations of either CycleGAN [45] or Pix2Pix [46], which are approaches for training a neural network for the task of image 2 image translation. Recently however, denoising diffusion models are becoming more and more popular for the task of image 2 image translation, and it would not be a surprise to see future renderers incorporate the power of these generative models. As the main contribution of this paper is a novel landmark generation module for the task of overdubbing, no further analysis is carried out on these modules as they fall outside the scope of this work.

B. 3D MODEL BASED METHODS

Even though the approach presented in this paper is a landmark based one, it is worth briefly discussing 3D model-based ones [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. Most of these approaches follow the high-level pipeline denoted by Figure 2 above.

Monocular reconstruction is carried out on each frame of the target video, generating a 3D mesh for every frame. From these meshes, pose, facial expression, and geometry parameters are extracted. The target audio is passed through a specially designed ''audio to expression'' network, that can generate blend shape expression parameters from the audio directly. Finally, the newly generated expression parameters are combined with the pose and geometry parameters from the original video, in order to generate a new set of meshes. These are then rendered back into photo-realistic frames with the help of a neural rendering network.

C. IMAGE RECONSTRUCTION METHODS

As mentioned earlier, these are the approaches which use pure image reconstruction techniques and latent feature learning [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], where one can feed in a reference image and target audio clip into the network, and output a photo-realistic talking head video. Because these are typically ''End-to-End'' systems, they have several advantages over structural methods: all parameters can be trained under one loss, they are typically faster, and can be deployed with more ease on neural inference chips. The quality of the videos they produce however are not as good as structural methods, and produce a lot more artifacts, especially when dealing with more extreme head poses. While these methods are exciting and show promising results, it was chosen to use a structural-based method for this paper, as the quality of the final rendered video is significantly better, and more control over the contents of the generated video can be exerted.

III. METHODOLOGY

Following a detailed review of the current literature for automatic dubbing, a gap was identified that provides the basis for this work. Typically, in image-based networks, where the goal is to generate a photo-realistic talking head video from a single reference image and target audio, the common approach is to have one network focus on generating the lip movements, and another network for generating the rest of the facial movements such as head pose, jaw movements, etc. The outputs of both networks are then combined to generate the fully animated facial landmark sequence like in [22]. For video-based methods, the goal is to modify a reference video given a target audio. The approach generally involves generating the lip movements first, before refitting them onto a landmark sequence extracted from the reference video.

As this approach is a video modifying task for the purposes of automatic dubbing, it is proposed to discard the intermediate processing steps mentioned above entirely and train a network to generate audio driven moving lips that are in alignment with 3D head pose extracted from the reference video directly. An advantage to this is that the overall animation pipeline is faster, simpler, saves on compute, and lends itself better to real time applications. Secondly, due to the unique pre-processing approach employed before training, classic normalisation techniques for this task such as removing speaker identity and head motion are not used, allowing the training data to maintain its structural integrity, and therefore the network can learn to generate more accurate

and expressive lip movements. This is evidenced by the strong results obtained by the approach in this paper from the subjective user study carried out as part of this work, comparing the method presented in this paper, against other relevant landmark-based techniques from the literature.

Therefore, the contributions in this paper are twofold:

1) A novel LSTM-based pipeline is introduced, that takes as input a target speech clip along with pose and identity parameters extracted from a reference video. The network outputs a pose, and identity aware 3D facial landmark sequence with the lips synchronised to the target speech clip. This approach works in the 3D space and does not use a static face model to first generate the lip movements before retargeting them to a moving one, separating this work from other similar approaches such as [19], [22], and [20]. The model directly outputs lips synchronized to audio, that also follow the head pose and movement of the speaker, simplifying the overall pipeline.

2) A novel data augmentation method is introduced for the pipeline training task, increasing the number of usable audio/visual pairs during training from N pairs up to NxN pairs, allowing the network to better learn the relationship between audio, lip expression, and pose. More precisely, Procrustes alignment is used to take the lip movements corresponding to a given audio signal and apply them to N additional landmark sequences, essentially ending up with a dataset where every audio sequence has N associated landmark sequences, each with unique head pose and movement, but with the lips being synchronized to that respective audio sequence. This augmentation helps when training the network as not only does it provide additional unique data, it decorrelates the lip movement from the rest of the face. During early experiments it was noticed that prior to adding this augmentation that lip movements become strongly correlated with global facial motion and head pose. Extensive details are provided in the data augmentation section on how to implement this and why it is important to do so.

An objective study evaluating the accuracy of the landmarks generated by this method against its ground truth was carried out and compared to other approaches in the literature. Additionally, a subjective user study was also carried out testing the quality of pose-aware landmarks versus other approaches by asking a series of carefully thought questions for each landmark sequence tested. The results of these experiments show it is possible to generate accurate, pose-aware landmarks at inference that are superior than other relevant approaches which use a static face shape and that by simply using the Procrustes lip augmentation at train time, one can generate accurate pose-aware landmarks using any existing method or architecture. Details on these experiments are provided in the results section.

To summarize, this work presents an automatic facial dubbing network that takes in a target speech audio and a reference facial landmark sequence as input. The network modifies the lip displacements of the reference landmark sequence in order to produce a new sequence whose mouth movements are correctly aligned with the speech audio while maintaining the original head movements and poses of the reference video. This is done to keep the actor's performance as close to the original as possible, and not to take away from its quality in any way. See Figure 3 below as it depicts a high-level overview of the network architecture.

A. DATA PROCESSING

1) DATA-SET SELECTION

While the end of goal of any automatic dubbing pipeline is for it to be subject / speaker independent, for the purposes of this paper a single speaker dataset was chosen to establish a proof of concept and determine the different elements of the training pipeline. Therefore, the Obama Weekly Address [47] data-set was chosen. A collection of nearly 300 frontal full-face videos of President Barrack Obama, consisting of over 18 hours of audio-visual content. This dataset was selected for the following reasons:

1) It contains high quality audio, available at a frequency of 48KHz to go with video available at several different resolutions. For the task at hand, a video resolution of 720p was chosen.

2) In most of the videos, President Obama is the only speaker on video, making it very easy to isolate his facial region using an off the shelf face detector, and extract his 3D facial landmark co-ordinates.

3) President Obama is an ideal subject, as in his weekly address speeches he always faces the camera, speaks clearly, and while there is a large amount of variation in the head pose, there are not many extremes.

The native frame rate of the dataset is 29.97 FPS. For the experiments in this paper, the videos were down sampled to 25FPS as it made aligning each frame of audio with its corresponding video frame a much easier task and ensured that no audio information would be lost, i.e., with a frame rate of 25fps, each frame in the video would have an associated audio sequence of 40ms. For training of the network, most of the videos in the dataset were used, with a train/validation/test split of 85/10/5 percent maintained. Lists of the names and indexes of the videos, as well as pre-processing code are available on the project GitHub page, which will be made openly available to the public with the paper.

2) LANDMARK EXTRACTION

Initially, an off-the-shelf facial landmark extractor provided by [48] was employed to extract 68 3D facial landmarks from the individual frames from the videos in the dataset. Unfortunately the quality of the predicted landmarks from this library was found to be highly inconsistent, and to contain a lot of global jitter that had to be eliminated using smoothing techniques. It was found however that even small amounts of smoothing caused the landmarks to lose fine details in the lip motion, reducing the overall quality of the ground truth which ultimately affected the network's ability to generate accurate lip motions. Due to this, it was decided to use the 468 key

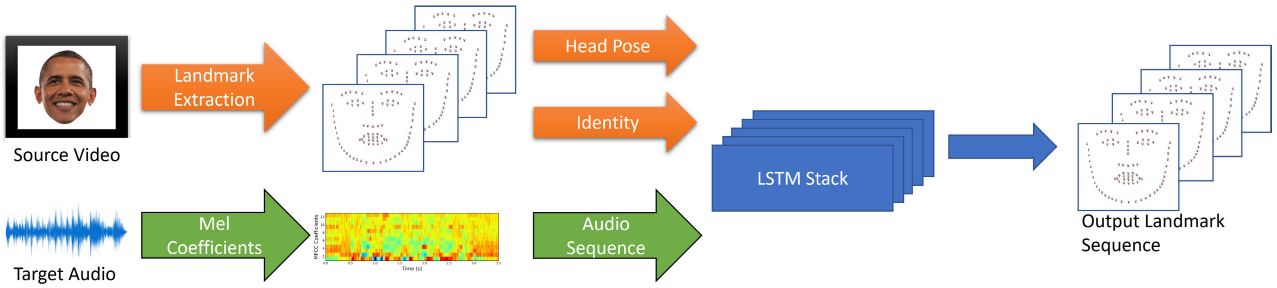


FIGURE 3. High level overview of architecture.

point face mesh extraction algorithm provided by Google's MediaPipe library [49]. Compared to [48], the 3D landmarks were near perfect, and more importantly, had virtually no global jitter present. For the sake of simplicity, 68 of the 468 keypoints that best resembled the landmarks returned by the traditional 68 keypoint dlib extractor were chosen to use as ground truth for training.

The landmark extraction process is very simple. First, the videos are processed with FFMPEG to get rid of any thumbnails at the start of the video or blank frames at the end. Second, individual frames are extracted from the training videos using the Open CV python library. Finally, the 468 3D landmark coordinates are extracted using the media pipe face mesh extractor, before selecting 68 custom keypoints that best resemble the traditional DLIB extractor to use for training the network. This process is done for every video in the dataset, with the code to do so available on the project GitHub page. To prepare the data for training, the landmark frames extracted from each video are combined such that a matrix of shape [N, 68, 3] is created for each video, where N is the total number of frames in that particular video.

Once the 3D facial landmarks have been extracted from every frame in every video, the next step is to normalise all the landmarks, and then apply a smoothing filter to get rid of any remaining jitter present. Normalisation is done by scaling the width of the face and centering the landmarks at the zero point like in [22]. The Savitzky-Golay filter is then used to smooth out the remaining jitter.

3) AUDIO FEATURE EXTRACTION

Once the videos are processed and the landmarks are extracted, the next step is to prepare the audio for training. The audio being used as part of the training set is single channel, has a sampling rate of 48000 Hz and is stored as a WAV file. Remember that since the framerate in the training videos is 25FPS, each frame covers 0.04 seconds of audio information.

The chosen audio features which are to be fed into the neural network are known as Mel Coefficients. These are state of the art features used in many related applications, most commonly in automatic speaker/speech recognition tasks. Reference [50] provide an in-depth explanation of what they are and how they are computed.

The audio signal is framed into 40ms frames, and various experiments were carried out training the network with a range of hop lengths starting from a hop length of size 1920 (no overlapping frames), to 960, to 480, ensuring various degrees of overlap between audio frames. It was decided to not use overlapping frames as no visible difference was noticed in the accuracy of the predicted landmarks against the original. A mel-filterbank of size 80 was also chosen. Therefore, for a 1 second audio sequence, the resulting feature matrix would have shape (25,80).

4) ALIGNING AUDIO WITH LANDMARKS

Now that the audio and landmark features are ready, the next step is to pair them together in preparation for training. For a given video V that contains T number of frames is depicted as V_T. Additionally, for the corresponding audio sequence A, which contains T-1 audio frames, is depicted as $A_{(T-1)}$. Notice that there is one extra video frame at the start of every sequence which is discarded from the audio/landmark pair. This is done as it is assumed that the audio preceding the frame influences it, therefore there is no need to keep the first frame in the sequence as it has no audio associated with it. Note that this assumption is made as the data is being fed into an lstm as a sequence, therefore the network has knowledge of past and future frames. Had we been using an architecture that would generate the output frame by frame, we would need to expand the audio window to cover future frames too. This is to ensure that facial movements caused by plosive sounds would be correctly learnt. The first frame in the video is instead saved as a separate entity from which the identity parameter is extracted for its associated sequence.

The final step is to combine these audio/visual frames into a sequence of 100 pairs for training. 100 is chosen as it is equivalent to 4 seconds worth of audio/visual content (25 fps $\times$ 4). This was a simple design choice influenced by the memory constraints of the available GPU. Please see Figure 4 below for a visual description of the alignment process. Note how the first frames in each of the 4 second sequences are discarded as explained above.

B. DATA AUGMENTATION

1) PROCRUSTES LIP AUGMENTATION

In this section "Procrustes Lip Augmentation" is introduced, a novel augmentation technique designed to increase the

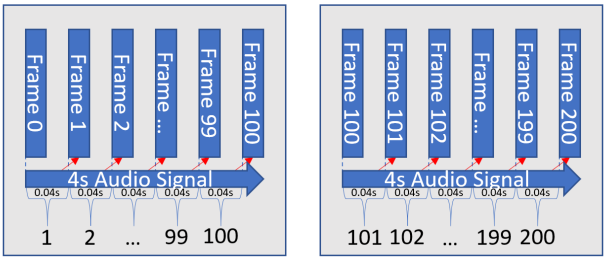


FIGURE 4. Landmark frame / audio sequence pairing process.

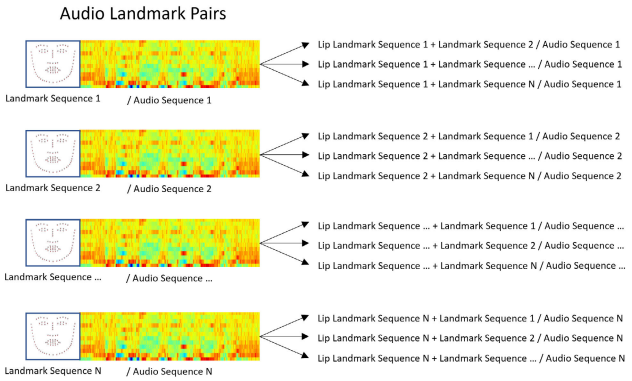


FIGURE 5. Visualisation of the Procrustes lip augmentation process.

number of usable audio-visual pairs available during training from N pairs up to NxN pairs, as well as decouple the relationship between the movement of a person's lips, and the direction, pose, and movement of their face.

Assuming one has a number of aligned audio/landmark sequences denoted as $A_0/L_0 \rightarrow A_N/L_N$ where N denotes the total number of sequences. For a given audio sequence A_0, it's associated lip landmarks are extracted from the overall landmark sequence L_0, and inserted into every other landmark sequence in the dataset using Procrustes Analysis. Through this, one can obtain N sequences of landmarks, where the lip movements are synchronized to the speech from A_0, while the head poses, and head movement are all unique. By doing this, one can successfully de-correlate the relationship between head pose, and lip movement. The following steps depict the process:

2) RATIONALE FOR LIP AUGMENTATION

The technique evolved from some initial experiments, where the model was being trained on speech from a single speaker, and having it output the aligned animated facial landmarks. In this initial training experiment, 4 second long sequences of audio combined with a head position vector at each frame were fed to the network, where the role of the position vector was to provide information about the head pose to the network. The intuition was that the network would take these inputs and use them to output the new pose aware facial landmark co-ordinates, with the lips being synchronised to the audio. The idea was that the audio would drive the movement of the lips, while the position sequence would tell the network the direction in which the head was facing, and generate the position of the lips on the face accordingly.

Rather than having the desired effect of outputting pose aware facial landmarks, the network ended up treating the audio portion of the input as noise, and completely ignoring it. Instead, the network learned how to generate accurate lip movements from the head position sequence alone. A number of tests were carried out to confirm this, specifically silence was fed into the network, along with a variety of head position sequences to test whether the network would still generate lip motion. The tests indicated that the network was ignoring the audio portion of the input entirely, as in each of the tests with silence, the generated lips would still be moving. Therefore it was concluded that there was a strong correlation between the movement the speakers lips in the dataset, and their head pose at any given frame.

This phenomenon led to the realization that in order to train a network to generate audio driven lip landmarks, it is crucial to de-correlate the relationship between the motion of the lips, and the head pose / general movement of the face. It is for this very reason that most approaches in the literature employ a static face mesh during training as it allows their models to learn the movement of the lips with respect to the audio, without having to worry about other aspects like head pose and facial movement. As the purpose of this work is to output pose aware moving lips, a workaround for this issue was necessary.

Initially it was believed that the model was overfitting on the single speakaer dataset, and introduced a multispeaker dataset during training to try and alleviate this issue. Despite this, the network continued to treat the audio portion of the input as noise, learning the lip movement from the head pose sequence alone. It was at this point that the idea to use the "Procrustes lip augmentation" came about. The augmentation had the desired effect, successfully de-correlating the relationship between the head pose, and lip movement in the training data set. This allowed the network to learn to output 3D facial landmark sequences, with the head pose controlled by the pose sequence extracted from a reference video, and the lip movement synchronised to and driven by the target audio. To replicate this augmentation, please see the steps below:

3) STEPS FOR PROCRUSTES LIP AUGMENTATION

1) Process the whole dataset as described in section 3.1, such that you have Audio/Landmark Sequence pairs ready.
2) "Procrustes analysis determines a linear transformation (translation, reflection, orthogonal rotation and scaling) of the points in Y to best conform them to the points in matrix X, using the sum of squared errors as the goodness of fit criterion" [51]. For a given audio/landmark pair, A_0/L_0, take L_0 and run Procrustes analysis against sequence L_1.

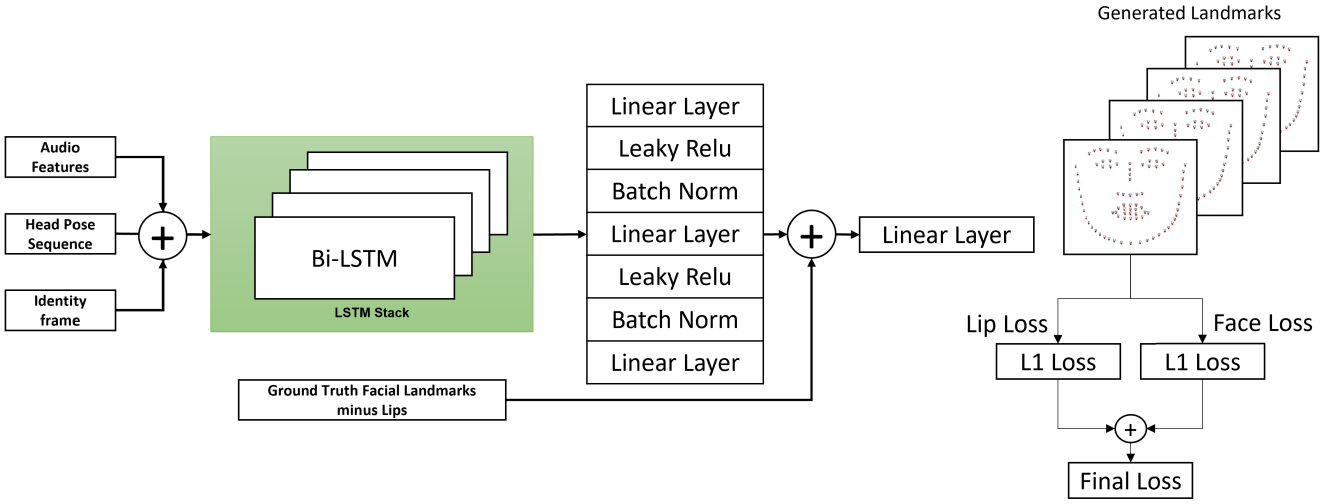


FIGURE 6. High level overview of model architecture.

TABLE 1. Detailed breakdown of the model layers displayed in figure 6 above. The input and output shapes, along with any relevant hyperparameters are included.

Layer Name	Input Shape	Output Shape	Number of Layers	Other Hyperparams
Bidirectional LSTM	89	256	4	Dropout = 0.5
Fully Connected Block:	-	-	-	-
Linear Layer	256	256	1	Bias = True
Batch Normalisation Layer	256	256	1	-
Leaky ReLu Layer	256	256	1	Negative Slope = 0.2
Linear Layer	256	128	1	Bias = True
Dropout	-	-	-	p = 0.5
Batch Normalisation Layer	128	128	1	-
Leaky ReLu Layer	128	128	1	Negative Slope = 0.2
Linear Layer	128	60	1	Bias = True

3) You will have to do it frame by frame. L_0 is Y, and L_1 is X. What you obtain is the conformed sequence $\hat{L}_0$
4) Isolate the lip landmark positions from $\hat{L}_0$ and use them to replace the lip landmark positions in L_1.
5) Repeat steps 2 and 3 for the rest of the landmark sequences in your dataset such that you end up with L_N modified sequences that are synchronized to A_0.
6) Repeat the steps above with the rest of the audio sequences in your dataset $A_{1\rightarrow N}$

Realistically though, one cannot do this for every single audio sample in the dataset as the processing time would take too long. Instead, for every audio sample, 10 random landmark sequences were chosen to do the Procrustes lip augmentation for, increasing the size of the training dataset by 10 times. The number 10 was chosen as no noticeable improvements in the accuracy of the network were discovered by increasing this number further. In fact, even applying the augmentation to 5 landmark sequences for every audio clip was found to be more than enough to de-correlate the audio from the head pose. Please see figure 5 for a visual representation of this process.

C. NETWORK ARCHITECTURE AND TRAINING SET UP

In this section the network architecture, and training set up of the work in this paper is discussed. With a focus on the choice of model, hyper-parameters, and rationale behind certain design choices.

1) NETWORK ARCHITECTURE

The network is a very simple LSTM-based neural network, that takes speech audio features as input combined with a head pose sequence and identity embedding. The network is trained to output the pose-aware facial landmark co-ordinates. This is depicted by the architecture diagram presented in Figure 6.

The audio features are sequences of Mel Coefficients spanning 4 seconds of audio each, as described in section 3.1. They have a shape of (99,80). The head pose sequence is extracted from each frame of the corresponding landmark sequence. For each frame, the ''pose'' is computed from 3 co-ordinates associated with the nose on the face. In total for a 4-second-long sequence, 99 such head pose embeddings are obtained, having a shape of (99,3,3). The head pose sequence array is then flattened, and concatenated with the mel coefficients array, ending up with a new training feature of shape (99,89). Recall that the first landmark frame in the sequence is removed as it has no equivalent audio information. This frame is saved, and from it the identity parameter is extracted by passing the landmarks extracted from the frame through 3 linear layers, reshaping it to be of size (1,89). This feature

is then inserted at the beginning of the training array, ending up with a final feature shape of (100,89).

The input is then passed through a stack of 4 bidirectional LSTMs with an input size of 89, and hidden size of 128. The output of the LSTMs then passes through a linear layer of in size 256, and out size 256. This then passes through a batch normalization layer followed by a leaky ReLU layer with a 0.2 slope coefficient. Another linear layer then takes the embedding of size 256 as input, and outputs one of size 128. Followed by a dropout layer, and another batch norm and leaky ReLU. Finally, one last linear layer of in size 128, outputs an embedding of size 60. This is a flattened set of 20 lip co-ordinates. Please see table 1 for a summary of all layer parameters.

Next, the original landmark sequence minus the lips is concatenated with the newly generated lip co-ordinates. This passes through a final linear layer of in size 204, out size 204 to smooth out any jitter. The output is our new set of generated landmarks for the given audio sequence.

For training the network, L1 loss is chosen as the loss function combined with the ADAM optimizer. Two losses are calculated, a lip loss, and a face loss. The lip loss simply takes the generated lips and compares them to the original lips, while the face loss takes the entire set of landmarks and compares them against the original. The lip loss is weighted 90 percent, while the face loss is given a weight of 10 percent.

2) TRAINING SET UP

The network is trained using a 3070-laptop edition GPU. The training data is prepared and extracted from 200 videos of the Obama Weekly Address dataset. The data is augmented as described in section 3.1, for every audio sequence, 10 random landmark sequences were chosen, and modified their lips such that they would be synchronised for the given audio. Ending up with 10 sequences of unique head motion per audio sequence. The network is trained on the augmented dataset for approximately 12 hours with a learning rate of 0.001 and the ADAM optimizer.The batch size is set to 512, and the Audio/Landmark sequences are shuffled for training.

IV. EXPERIMENTS AND RESULTS

In this section, the experiments and results of this paper are presented and discussed. The results in this work are subjectively compared to the results obtained by works presented in [22] and [20] as these are the methods most relevant to the one in this paper. It was attempted to also compare the model to the approach taken by [17] however the authors have not made the code necessary for this available. Additionally, an objective comparison is also provided between the generated landmarks of this paper versus the ground truth, and those of [19], [22], and [20] and their respective ground truth data. Note that both [19] and [20] use the same approach for generating landmarks. Because this work focuses on the landmark generation aspect of the automatic dubbing pipeline, the evaluation is carried out on the generated landmarks. Sample video renderings that are generated using landmarks extracted from the approach presented in this paper are provided as a proof of concept however a dedicated renderer to transform the 3D landmarks back to 2D RGB frames has not been trained, as that falls outside the scope of this paper.

TABLE 2. Mean opinion score per question.

Models	Q1↑	Q2↑	Q3↑	Q4↑	Q5↓
Ground Truth	3.975	3.839	3.961	3.836	2.554
ATVG Net [20]	2.607	2.982	2.454	2.475	3.514
MakeItTalk [22]	2.65	2.814	2.954	2.564	3.843
Proposed Approach	4.018	3.986	4.029	3.929	2.554

A. SUBJECTIVE USER STUDY

A subjective user study was carried out, evaluating the quality of the landmarks generated by the work in this paper, the work in [22], and the work in [20]. Ten different videos of President Barrack Obama speaking were evaluated per model, 30 (3 models $\times$ 10) videos in total. Each video had length 16 seconds. Additionally, ten ground truth videos were also evaluated as part of the study to be used as a baseline. In total, 28 subjects participated, evaluating 40 videos each. Note that the subjects were asked to evaluate videos produced using landmarks, and not the RGB frames. The scale of the study was kept small as the goal was to show that generating highly accurate, pose-aware landmarks at inference is possible, and that the accuracy of the generated lip movements using the method outlined in this paper is comparable to other ''static'' face based methods.

Subjects were asked to watch each of the 40 videos in random order, and to answer 5 questions per video to evaluate it. The subjects had a choice of 5 answers per question, which were ''Strongly Disagree'', ''Disagree'', ''Neutral'', ''Agree'', and ''Strongly Agree''. The subjects were not told which approaches were used to generate the particular video they were evaluating, nor were they told whether the video came from the generated or ground truth set. Table 2 contains a summary of the results, showing the mean score each model obtained per question while figure 7 contains a more detailed breakdown for each individual question. Note that the questions asked are listed above their respective tables. From these results it is clear that the approach presented in this paper produces a model capable of generating audio driven pose-aware landmarks that are near indistinguishable from the ground truth landmarks extracted directly from video. Readers are encouraged to view the generated videos provided in the supplementary materials section to see the accuracy of the model.

B. OBJECTIVE STUDY

Evaluating the the predicted landmarks in an objective manner is a non-trivial task. Distance based metrics are by far the most popular method of evaluating the predicted landmarks against their ground truth, and some type of a distance metric (usually L1/L2 distance) is often used as the loss function during the training phase. As part of this work, an objective study is carried out using the distance-based

Q1: The motion in the video was realistic overall (lips, head motion, etc).

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	3.975	0.092	3.787	4.163
ATVG_Net	2.607	0.165	2.268	2.947
MakeItTalk	2.65	0.162	2.318	2.982
Proposed Approach	4.018	0.099	3.815	4.221

Q2: The lip motion was synchronised well with the audio.

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	3.839	0.097	3.64	4.039
ATVG_Net	2.982	0.127	2.721	3.243
MakeItTalk	2.814	0.135	2.537	3.091
Proposed Approach	3.986	0.082	3.818	4.153

Q3: The head motion was synchronised well with the audio.

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	3.961	0.089	3.778	4.144
ATVG_Net	2.454	0.168	2.108	2.799
MakeItTalk	2.954	0.16	2.624	3.283
Proposed Approach	4.029	0.076	3.873	4.184

Q4: The motion appeared natural.

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	3.836	0.097	3.638	4.034
ATVG_Net	2.475	0.163	2.14	2.81
MakeItTalk	2.564	0.165	2.225	2.904
Proposed Approach	3.929	0.094	3.735	4.122

Q5: There were artefacts (distortions) in the motion.

Approach	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Ground Truth	2.554	0.157	2.232	2.875
ATVG_Net	3.514	0.117	3.273	3.755
MakeItTalk	3.843	0.149	3.537	4.149
Proposed Approach	2.554	0.141	2.263	2.844

FIGURE 7. Estimated marginal means calculated for each question the subjects answered.

metrics described by [17], comparing the accuracy in the predicted landmarks from a range of models against their respective ground truths. The ground truth landmarks associated with each of the models were extracted from their respective test sets, and pre-processed in accordance with the instructions provided by their respective GitHub pages, and papers. The landmark distance, and landmark velocity difference [20], [22] functions are used to evaluate the predicted landmarks against their ground truths. The results of these evaluations are provided for in figure 9. Like in [17], the LD and LVD functions are used on the mouth and face area separately. This is denoted by M-LVD, M-LD, F-LVD, and F-LD respectively. Note that F-LD and F-LVD are very low for this paper compared to other approaches because the

TABLE 3. Objective evaluation results against GT.

Models	M-LD↓	M-LVD↓	F-LD↓	F-LVD↓
ATVG Net [20]	7.111%	0.947%	7.149%	0.719%
MakeItTalk [22]	8.492%	0.391%	8.896%	0.507%
Proposed Approach	3.042%	0.332%	0.178%	0.001%

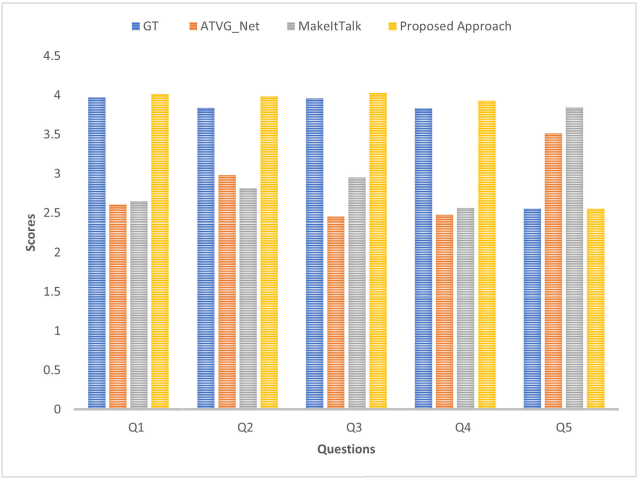


FIGURE 8. Plot of mean scores each model obtained per question.

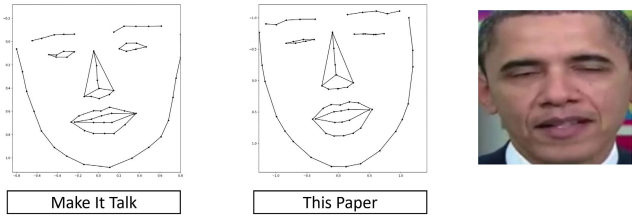


FIGURE 9. Comparison of ground truth landmarks extracted from the same frame.

model is trained with knowledge of what the face looks like, and the direction in which the head is facing.

C. INTERPRETING RESULTS

Both the subjective and objective studies carried out as part of this work show that the approach presented in this paper for generating 3D pose-aware landmarks is a feasible one for generating accurate, and expressive talking head landmarks. It is indeed possible to generate high quality, pose-aware landmarks at inference, without a suffering losses in the quality of the lip synchronization. Based on the subjective results it is clear that subjects preferred talking heads that had identity, and pose information. While the approach presented in this paper slightly outperforms the ground truth in most of the question categories, this can be simply attributed to the very high similarity between the ground truth landmarks and generated ones. A common piece of feedback from subjects who did the study was that they were confused why they were shown two of the same video (recall that the information that one was ground truth and one was generated was not revealed).

Additionally, it can be seen that the approach presented in [22] outperforms [20] in categories related to overall

motion, motion naturalness, and head motion, while scoring slightly worse on lip/audio synchronization and motion artefacts. This tracks well as [22] approach is capable of generating realistic head motion for the landmarks, and it can be seen from this study that subjects noticed and preferred that over [20].

V. CONCLUSION

The goal of this paper was to introduce an approach for generating 3D pose and identity aware talking head landmarks given a source video and driving speech signal for the task of automatic video dubbing. It is shown throughout the paper that this is quite feasible to do via a novel data augmentation technique, and that subjects preferred landmarks generated by this approach over other, existing approaches such as [20], [22], and [19]. A number of key insights were gained by conducting this work:

1) Generating 3D pose aware landmarks is possible, and can be easily achieved by de-correlating the relationship between the lips, jaw, and global head movement through the Procrustes lip augmentation that is proposed in this work.

2) The quality of the ground truth data is much more important over the choice of model for learning the relationship between audio and lip movement. Oftentimes noisy data needs smoothing, and smoothing leads to losing valuable lip motion, therefore the model wont be able to learn anything meaningful from audio. This can best be seen from the results of the subjective study where models trained with inferior ground truth scored poorly on metrics such as naturalness and motion distortions.

3) By carrying out a subjective study, and surveying 28 users, it was shown how important the inclusion of head movement information is when evaluating the quality of talking head landmarks.

As part of this work, all data-sets, code, and trained model weights will be made available to the community.

A. FUTURE WORK

This research has opened up a number of potential avenues for future work. At the forefront of these, is the idea to develop a generalised pose-aware model with the capability to few-shot learn individual speaking styles. Over the course of this work, it was discovered that when training a landmark prediction network on a single speaker, the network was robust to generating landmarks from a wide variety of speakers. Regardless of what speech was being input to the network, it was observed that the network would always generate accurate landmarks but in the speaking *style* of President Obama. This indicates that it may be possible to train a generalised model and teach it via few-shot learning techniques to output landmarks in a specific speaker style given a very small amount of data of that speaker.

A dedicated neural renderer for the task of landmark based automatic dubbing is also in the works. Sample renderings were generated using the pretrained model provided by [22] as a proof of concept, however it does not handle extreme variations in the head pose very well as it is an image-based renderer. These can be seen in the supplemental videos section. Training of a video-based renderer is necessary to generate the best possible results. Recent advances in generative neural networks related to diffusion models seem like a promising avenue to explore.

Additionally there is still room to improve the lip landmark generation, increasing its robustness to unseen speakers via deep-learning based audio augmentation techniques such as voice cloning, and synthetic speech generation, as well as more classical approaches like pitch variation, time warping, or noise addition.

B. LIMITATIONS

There are several limitations when generating pose aware landmarks using the method presented in this paper.

1) The network does not generate realistic jaw movements from audio. Due to the nature of the data augmentation (de-correlating lip motion from jaw/head movement), the network is not able to learn to also generate the corresponding jaw movement from the audio. This limitation can be overcome by computing the distance between the upper lip, and lower lip, and raising/lowering the position of the jaw by this amount via a simple linear equation. Alternatively, a very simple network can be trained to solve this, consisting of just a couple of LSTM layers as there is a very direct correlation between lip and jaw movement that can be learnt.

2) Throughout this work it is shown that head pose is related to audio, and a method is demonstrated to decouple this relationship. Due to this, the approach in this paper is not a suitable one for audio-driven video generation. Rather than generating talking head videos from scratch, the proposed network learns to modify an existing video, keeping the original headpose but changing the lip content in response to a new audio signal. This is ideal for the task of dubbing, as it is assumed that the speech content and emotion of the dubbed speech is similar to that of the original. Therefore it is desired that the performance of the actor in the generated video is kept as close to the original performance as possible, including the head movements. However, this is a limitation, because when inputting new speech content that doesn't necessarily match the original headpose, such as silence, the resulting output will contain the original head motion, but with the lips firmly shut. This may lead to the user perceiving the resulting video as being ''unnatural'', however more study in this direction is needed.

3) The approach presented in this paper is a single speaker approach. Because the network was trained using videos and audio from a single speaker (President Barrack Obama), it should not perform as well when exposed to audio from different speakers. That being said, the network is very robust, generating accurate and realistic landmarks from speech coming from a wide variety of speakers who were unseen to the network. Instead, it was observed that the speaking ''style'' of the output landmarks was very similar to that of President Obama regardless of the identity of the input

speech. It is possible to extend this work to be a multi-speaker network by training the network with data processed using the same techniques as employed by the works presented by [22] and [20] combined with the Procrustes lip augmentation.

4) Distance-based metrics are not that useful when attempting to judge the quality of landmarks generated across different models in the literature, especially when one tries to make a direct comparison between said models. For example, consider model A, trained using landmarks extracted, and normalised using pre-processing method A. The quality of the landmarks generated by model A are only as good as the ground truth model A was trained on. Furthermore, any distance metric used for evaluating the landmarks, will be calculated using the predicted landmarks, and it's associated ground truth, therefore one cannot directly compare the landmarks from model A and model B with distance based metrics as they are both likely to have different methods for extracting their ground truths. See figure 8 to see just how different the landmarks extracted from the same frame can be. Due to the reasons outlined above, it is entirely possible that in a comparison between two models, A, and B, where model A has inferior ground truth to B due to variations in the landmark extraction process, model A could report better scores than B even though B may look visually better. Despite this, it is still very useful to provide distance based comparisons between other similar models in the literature, and their respective ground truths, as it helps one gain a rough idea regarding the quality of their generated landmarks with respect to other approaches.

5) As this is an approach towards generating audio driven pose-aware landmarks, rendering the landmarks falls outside the scope of this work. That being said, example renderings of the landmarks generated by this approach using the pretrained renderer from [22] are provided in the supplementary videos section. These videos are there as a proof of concept, showing that one can render high quality videos from the 3D pose-aware landmarks presented in this paper.

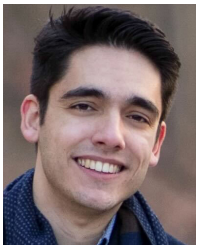
C. FINAL REMARKS

While automated dubbing has implications for deep-fakes, it is becoming a reality and the benefits for making entertainment more readily available to a wider and more global audience is important - this doesn't just mean English to other languages - it can also mean content in low-resource languages dubbed back into more realistic English! Major streaming companies already have a lot of non-English content so this is important for the further democratisation of content.

REFERENCES

[1] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy, "Everybody's Talkin': Let me talk as you want," 2020, *arXiv:2001.05201*.
[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.
[3] C. Zhang, S. Ni, Z. Fan, H. Li, M. Zeng, M. Budagavi, and X. Guo, "3D talking face with personalized pose dynamics," *IEEE Trans. Vis. Comput. Graph.*, early access, Oct. 4, 2021, doi: 10.1109/TVCG.2021.3117484.
[4] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, Jul. 2017.
[5] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, "Audio-driven talking face video generation with learning-based personalized head pose," 2020, *arXiv:2002.10137*.
[6] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," 2019, *arXiv:1905.03079*.
[7] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 3867–3876.
[8] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 3660–3669.
[9] H. Wu, J. Jia, H. Wang, Y. Dou, C. Duan, and Q. Deng, "Imitating arbitrary talking style for realistic audio-driven talking face synthesis," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1478–1486.
[10] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler, "LipSync3D: Data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 2754–2763.
[11] A. Richard, M. Zollhofer, and Y. Wen, "MeshTalk: 3D face animation from speech using cross-modality disentanglement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 1173–1182.
[12] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," 2019, *arXiv:1912.05566*.
[13] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu, "Photorealistic audio-driven video portraits," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 12, pp. 3457–3466, Dec. 2020.
[14] L. Song, B. Liu, G. Yin, X. Dong, Y. Zhang, and J.-X. Bai, "TACR-Net: Editing on deep video and voice portraits," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 478–486.
[15] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," 2020, *arXiv:2007.08547*.
[16] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–11, Jul. 2017.
[17] X. Ji, H. Zhou, K. Wang, W. Wu, C. Change Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," 2021, *arXiv:2104.07452*.
[18] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2Head: Audio-driven one-shot talking-head generation with natural head motion," 2021, *arXiv:2107.09293*.
[19] S. E. Eskimez, R. K Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," 2018, *arXiv:1803.09803*.
[20] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7824–7833.
[21] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: Real-time photorealistic talking-head animation," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–17, Dec. 2021.
[22] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: Speaker-aware talking-head animation," 2020, *arXiv:2004.12992*.
[23] D. Aneja and W. Li, "Real-time lip sync for live 2D animation," 2019, *arXiv:1910.08685*.
[24] S. Biswas, S. Sinha, D. Das, and B. Bhowmick, "Realistic talking face animation with speech-induced head motion," in *Proc. 12th Indian Conf. Comput. Vis., Graph. Image Process.*, Jodhpur, India, Dec. 2021, pp. 1–9.
[25] W. Wang, Y. Wang, J. Sun, Q. Liu, J. Liang, and T. Li, "Speech driven talking head generation via attentional landmarks based representation," in *Proc. Interspeech*, Oct. 2020, pp. 1326–1330.
[26] N. Sadoughi and C. Busso, "Speech-driven expressive talking lips with conditional sequential generative adversarial networks," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1031–1044, Oct. 2021.
[27] D. Das, S. Biswas, S. Sinha, and B. Bhowmick, "Speech-driven facial animation using cascaded GANs for learning of motion and texture," in *Computer Vision—ECCV 2020*, vol. 12375, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 408–424.
[28] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.
[29] T. Xie, L. Liao, C. Bi, B. Tang, X. Yin, J. Yang, M. Wang, J. Yao, Y. Zhang, and Z. Ma, "Towards realistic visual dubbing with heterogeneous sources," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1739–1747.

[30] R. Zhao, T. Wu, and G. Guo, "Sparse to dense motion transfer for face image animation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 1991–2000.
[31] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," 2020, *arXiv:2008.10010*.
[32] G. Mittal and B. Wang, "Animating face using disentangled audio representations," 2019, *arXiv:1910.00726*.
[33] H. Zhu, H. Huang, Y. Li, A. Zheng, and R. He, "Arbitrary talking face generation via attentional audio-visual coherence learning," 2018, *arXiv:1812.06589*.
[34] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "End-to-end generation of talking faces from noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1948–1952.
[35] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," 2018, *arXiv:1803.10404*.
[36] H. Zhou, Y. Sun, W. Wu, C. Change Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," 2021, *arXiv:2104.11116*.
[37] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," 2019, *arXiv:1906.06337*.
[38] N. Kumar, S. Goel, A. Narang, and M. Hasan, "Robust one shot audio to video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 3334–3343.
[39] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," 2018, *arXiv:1807.07860*.
[40] Y. Song, J. Zhu, D. Li, X. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," 2018, *arXiv:1804.04786*.
[41] J. Son Chung, A. Jamaludin, and A. Zisserman, "You said that?" 2017, *arXiv:1705.02966*.
[42] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "JALI: An animator-centric viseme model for expressive lip synchronization," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, Jul. 2016.
[43] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "VisemeNet: Audio-driven animator-centric speech animation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–10, Aug. 2018.
[44] Y. Guo, K. Chen, S. Liang, and Y.-J. Liu, "AD-NeRF: Audio driven neural radiance fields for talking head synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 5784–5794.
[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017, *arXiv:1703.10593*.
[46] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arXiv:1611.07004*.
[47] *Your Weekly Address*, The White House, Washington, DC, USA, May 2015.
[48] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1021–1030.
[49] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.
[50] L. Roberts, "Understanding the Mel spectrogram," Tech. Rep., Mar. 2020. [Online]. Available: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53
[51] *Procrustes Analysis—MATLAB Procrustes*. [Online]. Available: https://www.mathworks.com/help/stats/procrustes.html



DAN BIGIOI (Graduate Student Member, IEEE) received the bachelor's degree in electronic and computer engineering from the University of Galway, in 2020, where he is currently pursuing the Ph.D. degree, sponsored by D-REAL, SFI Centre for Research Training in Digitally Enhanced Reality. Upon graduating, he worked as a Research Assistant at the University of Galway studying text to speech and speaker recognition methods under the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project. His research interests include studying and implementing novel deep learning-based techniques for automatic speech dubbing and discovering new ways to process multi-modal audio/visual data.

HUGH JORDAN received the B.A.I. and M.A.I. degrees in computer engineering from Trinity College Dublin, in 2021. He is currently a Ph.D. Researcher with Trinity College Dublin and the SFI Centre for Research Training in Digitally-Enhanced Reality. His research interests include audio-driven facial animation for automatic dubbing and virtual humans.

RISHABH JAIN (Member, IEEE) received the B.Tech. degree in computer science and engineering from the Vellore Institute of Technology (VIT), in 2019, and the M.S. degree in data analytics from the National University of Ireland Galway (NUIG), in 2020, where he is currently pursuing the Ph.D. degree. He is also working as a Research Assistant at NUIG under Data-Center Audio/Visual Intelligence on-Device (DAVID) project. His research interests include machine learning and artificial intelligence, specifically in the domain of speech understanding, text-to-speech, speaker recognition, and automatic speech recognition.

RACHEL MCDONNELL is an Associate Professor of creative technologies at Trinity College Dublin, Ireland. Her research focuses on animation of virtual characters, using perception to both deepen our understanding of how virtual characters are perceived, and directly provide new algorithms and guidelines for industry developers on where to focus their efforts. She has published over 100 papers in conferences and journals in her field, including many top-tier publications at venues such as SIGGRAPH, Eurographics, and IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS. She is a regular member of many international program committees (including ACM SIGGRAPH and Eurographics). She serves as an Associate Editor for journals, such as *ACM Transactions on Applied Perception*, *Computers and Graphics*, and *Computer Graphics Forum*.

PETER CORCORAN (Fellow, IEEE) currently holds the Personal Chair in electronic engineering with the College of Science and Engineering, National University of Ireland Galway (NUIG). He was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He has more than 600 cited technical publications and patents, more than 120 peer-reviewed journal articles, 160 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is also a member of the IEEE Consumer Technology Society for more than 25 years and the Founding Editor of *IEEE Consumer Electronics Magazine*.

Appendix J

Synthetic Speaking Children – Why we Need Them and How to Make Them.

Authors: Muhammad Ali Farooq (MAF), Dan Bigioi (DB), Rishabh Jain (RJ), Wang Yao (WY), Mariam Yiwere (MY) and Peter Corcoran (PC)

Contribution Criterion	*Contribution Percentage*
Research Hypothesis	MAF: 40%, DB: 15%, RJ: 15%, WY: 10%, PC: 20%
Experiments and Implementation	MAF: 50%, DB: 20%, RJ: 20%, WY: 10%
Background	MAF: 30%, DB: 20%, RJ: 20%, WY: 20%, MY: 10%
Manuscript Preparation	MAF: 60%, DB: 5% , RJ: 15%, WY: 10%, MY: 5%, PC: 5%

Synthetic Speaking Children – Why We Need Them and How to Make Them

Muhammad Ali Farooq, Dan Bigioi, Rishabh Jain, Wang Yao, Mariam Yiwere, Peter Corcoran
College of Science and Engineering, University of Galway, Galway, Ireland
muhammadali.farooq@universityofgalway.ie, d.bigioi1@nuigalway.ie, rishabh.jain@universityofgalway.ie, w.yao2@universityofgalway.ie, mariam.yiwere@universityofgalway.ie, peter.corcoran@nuigalway.ie

2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) | 979-8-3503-2797-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/SpeD59241.2023.10314943

***Abstract*—Contemporary Human-Computer Interaction (HCI) research relies primarily on neural network models for machine vision and speech understanding of a system user. Such models require extensively annotated training datasets for optimal performance and when building interfaces for users from a vulnerable population such as young children, GDPR introduces significant complexities in data collection, management, and processing. Motivated by the training needs of an Edge-AI smart-toy platform this research explores the latest advances in generative neural technologies and provides a working proof-of-concept of a controllable data-generation pipeline for speech-driven facial training data at scale. In this context, we demonstrate how StyleGAN-2 can be fine-tuned to create a gender-balanced dataset of children's faces. This dataset includes a variety of controllable factors such as facial expressions, age variations, facial poses, and even speech-driven animations with realistic lip synchronization. By combining generative text-to-speech models for child voice synthesis and a 3D landmark-based talking heads pipeline, we can generate highly realistic, entirely synthetic, talking child video clips. These video clips can provide valuable, and controllable, synthetic training data for neural network models, bridging the gap when real data is scarce or restricted due to privacy regulations.**

Keywords*—*Synthetic Data, Talking Head Generation, Text to Speech Synthesis, Facial Image Generation, Low Resource Data

I. Introduction

In the dynamic landscape of human-centric machine vision and speech analysis, researchers are frequently confronted with substantial challenges stemming from GDPR guidelines. Contemporary research heavily relies on neural network models and the availability of extensive training datasets to attain optimal performance. However, when the research focus shifts to the development of Human-Computer Interaction (HCI) interfaces and necessitates data from vulnerable populations, particularly young children, to train Edge-AI HCI models, GDPR introduces a myriad of complexities associated with data collection, management, and processing. These complexities are particularly pronounced when dealing with real data involving children, where stringent privacy regulations come into play.

In response to these challenges, recent advancements in Generative Adversarial Networks (GANs) and other generative neural technologies have emerged as promising solutions for generating data at scale. In this context, this paper introduces an innovative approach that leverages the power of such technologies. Motivated by the need for training data for an Edge-AI-based smart-toy platform [1] we demonstrate the adaptability of StyleGAN-2 [2], [3], a state-of-the-art generative neural architecture, to craft a gender-balanced dataset of synthetic children's faces. This dataset offers nuanced control over various critical attributes, including facial expressions, age variations, [1] facial poses, and the synchronization of facial movements with speech-driven animations, culminating in a collection of strikingly realistic videos.

Going beyond visual representation, our exploration extends into the domain of voice generation. By incorporating techniques such as FastPitch, advanced voice augmentation, and generative text-to-speech models, we achieve the ability to synthesize authentic children's voices, replete with their distinctive qualities. These voices, when seamlessly integrated with our StyleGAN-2 framework and speech-driven neural lip synchronization models, empower us to create highly realistic, entirely synthetic talking child videos.

These synthetic videos represent a pragmatic solution to data scarcity or stringent privacy regulations like GDPR. In addition to their applications in research and development, these videos serve as valuable training data for neural network models in a practical use case, such as an Edge-AI smart-toy platform. In developing these tools our focus has been on scaling to enable controllable data generation at scale. Thus spoken phrases can be employed in combination with a set of synthetic voices and multiple seed faces to fine-tune the computer vision and automated speech recognition models that operate on the smart-toy platform. This is useful, for example, to test how well the smart toy can detect the emotional state of a child or respond to variations in the command set for an interactive play activity. Gathering such data from children and directing their responses in a controlled laboratory environment is both time-consuming and costly.

The rest of this paper is devoted to providing a detailed explanation of our synesthetic child media generation pipeline starting with Section II which covers the creation of synthetic face samples, Section III which covers the generation of synthetic voice samples with FastSpeech 2, and Section IV which details how videos are created using MakeItTalk. Section V offers details on our experimental setup, with Section VI concluding the contents of this work.

II. An Overview of ChildGAN

The first step includes generating large-scale synthetic child facial data using advanced data augmentation methods. This is achieved by fine-tuning StyleGAN2 [2] for generating photo-realistic child data samples. This new synthetic child dataset is referred to as ChildGAN [5].

A. Training Methodology

StyleGAN2 is fine-tuned by using a transfer learning-based methodology. Transfer learning on GANs is a powerful

technique, especially when there are limited amounts of data and computational resources. It allows us to leverage the knowledge and representations learned by pre-trained CNN models and adapt them for new tasks and domains. For this study, we have trained the adapted StyleGAN2 model using a synthetic child dataset as the seed data. The original seed data is taken from adult data samples and is transformed using various algorithms, including GANs and Android-based mobile apps that create a child facial image from an input adult image. The overall process involves the fine-tuning of the StyleGAN2 generator, and discriminator, by using an adversarial training methodology. The complete training method is detailed in the work of [5].

B. Dataset and Quality Checks

We assess the quality of the synthetic data by employing various computer vision methods, ensuring the excellence and detail of the facial features in the generated data. To accomplish this, we utilize a combination of qualitative and quantitative metrics. These metrics encompass essential tests, such as face localization and facial landmark detection using the DLIB framework. Another crucial assessment involves the computation of the cosine similarity index, facilitated by ArcFace [6], to measure the similarity in identity among synthetic child faces. Additionally, we validate the quality of the artificial child faces for downstream applications by conducting tests with child gender classifiers, allowing us to evaluate the performance of these classifiers on real data [5]. In Fig. 1, we present a visual representation illustrating synthetic facial data featuring both boys and girls, generated through the ChildGAN network.

Fig. 1. Four distinct child facial samples of boys and girls generated from ChildGAN by finetuning StyleGAN2 with latent space editing.

C. Facial Transformations and Tools

The rendered synthetic data is further transformed to incorporate various smart transformations that can be used for diversified real-world computer vision applications. This is achieved by using the latent space editing feature in StyleGAN2. These transformations include eye blinking effects, age progression, directional lighting conditions covering different facial angles, facial expressions, head pose variations, and lastly hair and skin tone digitization. The complete dataset along with pretrained models are open sourced which can be used for more extensive data generation, further experimental analysis, and other related downstream tasks. Fig. 2 shows two different smart data transformations done via latent space editing [8] and relighting [7] based deep learning networks.

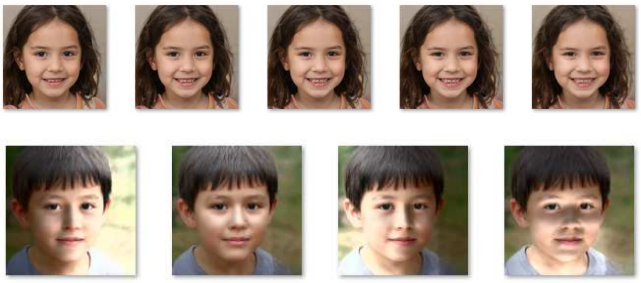


Fig. 2. Rendered child facial data with further advanced data augmentation results. The first row shows neutral to happy facial transformation on the generated girl subject and the second row depicts four directional lighting conditions embedded via [6] on the sample boy subject.

III. Synthesizing Children's Voices

Employing Cleese-based pitch augmentation [9], FastPitch TTS [10], and Tacotron2 TTS [11] for generating synthetic child speech holds immense potential for research applications. Cleese-based pitch augmentation allows for precise control over pitch contours, enabling the creation of highly realistic child voices with varying age and gender characteristics. FastPitch and Tacotron2, as cutting-edge TTS models, ensure the conversion of text into natural and expressive child-like speech. By leveraging synthetic data, researchers can conduct experiments without ethical concerns associated with using real child participants, while also ensuring reproducibility and standardization.

A. Text-to-Speech Child Voice Synthesis using Tacotron 2

The multi-speaker TTS model [12] consists of three distinct neural network models, each addressing a specific subtask: the Speaker Encoder for speaker verification, the Acoustic model for spectrogram synthesis, and the WaveRNN Vocoder for audio waveform generation.

The Speaker Encoder is trained using a combination of adult and child speech data from various datasets. It utilizes the Generalized End-to-End (GE2E) [13] loss to generate fixed-dimensional speaker embeddings. These embeddings enable the model to effectively distinguish between different speakers, allowing for better generalization across various voices. During training, complete utterances are segmented into partial utterances of 1.6 seconds, and the encoder is optimized over GE2E loss to ensure similar voices are mapped closer together in a latent space representation.

The Tacotron 2 Acoustic model [11], originally designed for single-speaker TTS, is adapted for multi-speaker functionality by incorporating the speaker embeddings alongside the text embeddings. The model is first trained with adult speech data and then fine-tuned with child speech data. The combination of speaker and text embeddings enhances the model's capability to generate spectrograms from input text conditioned on the specific speaker identity.

For audio waveform generation, the researchers employ the WaveRNN Vocoder [14], an improvement over the WaveNet model. WaveRNN is particularly chosen for its ability to perform sequential modeling of audio from mel-spectrograms. It utilizes a Gated Recurrent Unit (GRU) to

replace convolutions used in WaveNet [15], reducing sampling time while maintaining high output quality. The vocoder is trained on adult speech data and proves to be effective even with unseen speakers in multi-speaker models.

The proposed approach exhibited promising results in generating high-quality synthetic child voices which was verified using various subjective and objective evaluations.

B. Augmentation Techniques for Adult Voices

To generate synthetic child-like speech data from existing adult speech, a python-based sound manipulation toolkit known as Combinatorial Expressive Speech Engine (CLEESE) [9] is used to augment the adult speech data to make them closer to child voices. A d-vector based speaker encoder is used to compare the adult speaker embeddings to the mean child embedding to select the adult speakers most similar/proximate child speakers for the augmentation based on the cosine similarity metric. Specifically, the pitch and speaking rate of the selected adult speakers are raised and slowed down through the CLEESE pitch-shift and time-stretch transformations respectively, causing them to sound more child-like.

Objective and subjective (Mean Opinion Score (MOS)) evaluations performed showed that the Cleese-based augmentation approach successfully tuned the adult voices to sound child-like; however, due to the adult linguistic content and the absence of child-like prosodic features such as long pauses and "stammering", the augmented speech lacked the naturalness of real child speech. The evaluations also revealed that adult female speakers generally provided a better starting point for the augmentations as compared to adult male speakers. Overall, the average MOS score of 3.7 was reported for how convincing the augmented speech samples are as child speech and 4.6 for how intelligible the augmented speech is, for the best set of augmentation parameters. The work has been submitted to the IEEE ACCESS journal and is currently in the review process. In future work, we plan to improve the time-stretch transformation (speaking rate augmentation) in addition to modeling the child-like prosody and other paralinguistic features as part of the current augmentation approach; this is expected to improve the similarity between augmented speech and real child speech.

C. Using FastPitch to Synthesize Child Voices

A transfer learning pipeline is used for generating synthetic child voices using the Fastpitch TTS [10] model. The process involves pretraining the model with the LibriTTS [16] dataset, which includes diverse adult speech data. Then, the model is fine-tuned on a small subset of child speech data (MyST dataset [17]) to capture the acoustic properties and pitch contours specific to child speech. The finetuning pipeline is consistent with previous approaches using Tacotron 2 [11]. The vocoder used for generating high-quality speech waveforms is WaveGlow, which operates based on a generative flow-based model architecture. The WaveGlow [18] model is trained on the LibriTTS adult speech data and is employed as a universal vocoder for synthetic child voices.

Objective evaluations on the naturalness and intelligibility of the generated speech are conducted, comparing the Fastpitch model's performance with Tacotron 2 [11] for child speech synthesis. Moreover, speaker similarity verification using a pretrained speaker verification system shows that the synthetically generated child speech is close to real speech in terms of speaker similarity. This methodology successfully synthesizes realistic child voices, and the experimental results support the effectiveness of the Fastpitch model in generating high-quality synthetic child speech.

IV. Synthetic Talking Head Generation

Talking head generation presents a multitude of intricate challenges, including the precise synchronization of lip movements with speech, the maintenance of natural facial expressions throughout the animation process, and the overall cohesiveness of facial dynamics with the spoken content. Additionally, issues related to data quality, articulatory variation across different speakers, and achieving a high degree of realism in the generated faces are all formidable hurdles in this domain. These hurdles are further amplified when trying to generate synthetic child data, as children's speech and facial expressions exhibit unique characteristics and idiosyncrasies that demand specialized handling. Children's facial features and articulatory patterns differ significantly from those of adults, making it essential to tailor the synthesis process to capture these nuances accurately. Ensuring the generated child faces are both age-appropriate and realistic adds an extra layer of complexity to the task.

Existing talking head generation approaches often overlook these specific challenges, primarily because they are predominantly trained on adult data. Consequently, adapting such models for child-focused applications presents an open challenge in the field. The need to address the distinct nuances of child speech and facial expressions, while also maintaining the integrity of age-appropriate visual representations, underscores the gap that our research aims to highlight. With that in mind, it must be made clear that while our research does not explicitly tackle the unique issues associated with child-specific talking head generation, it serves as a foundational step toward exploring the potential of synthesizing child faces in this context.

In the current work, our focus is primarily on leveraging established techniques to create a synthetic dataset that can potentially catalyze further investigations and innovations in the field, paving the way for more specialized solutions to address the intricacies of child-focused talking head generation.

A. Rendering the Synthetic Talking Child Faces

MakeItTalk [4] is a structural-based talking head generation approach that works by generating a sequence of sparse 3D facial landmarks given a driving audio signal as input, followed by an image-to-image translation-based rendering step that generates realistic video frames from the landmark sequence and an input “seed” image. Two models are used to accomplish this, an audio-aware LSTM-based model for generating landmark sequences time-aligned to the input audio signal, and a Pix2Pix-based image rendering network.

In theory, any recent talking head generation model can be used for this step, however, we chose to use MakeItTalk [4] as it displayed remarkable robustness when exposed to synthetic voice data as input. We theorize that this can be attributed to how MakeItTalk handles audio input. Specifically, it adopts a process that disentangles the input speech into two distinct latent representations: a content embedding and a speaker ID embedding. This appears to bolster the model's adaptability and its ability to generate accurate and contextually relevant facial landmarks, making it an ideal choice for our use case.

In total, we generate and provide 20 synthetic child-speaking videos comprising of both boy's and girl's facial samples, each uniquely characterized by identities meticulously crafted through the ChildGAN model, and speech samples synthesized in accordance with the detailed process outlined in Section III. A high-level overview of this pipeline is depicted in Fig. 3.

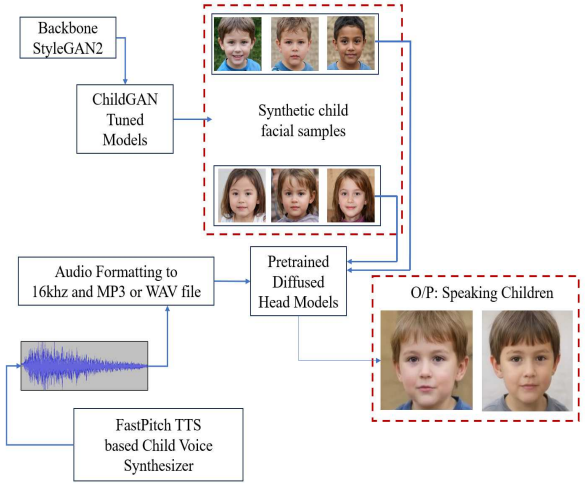


Fig. 3. Block diagram representing the pipeline adapted for generating 3D synthetic child speaking clips.

These synthetic videos display the capabilities of our synthesis pipeline. By offering this framework, we aim to facilitate a deeper understanding of child-like facial expressions and their correlation with speech, while also providing a valuable resource for researchers across various fields. Furthermore, we make all code, and scripts associated with this research publicly available.

B. *Evaluating Synthetic Videos*

We have abstained from presenting a formal evaluation of the synthetic videos we provide in this study. The rationale behind this decision is rooted in the fact that the videos we generate using our pipeline share a similar nature with those produced by MakeItTalk, which serves as a fundamental component of our methodology. Since MakeItTalk is a well-established framework for talking head generation with a recognized set of evaluation metrics and benchmarks, it offers a reliable reference point for the assessment of synthetic videos generated through our approach.

For a comprehensive and in-depth analysis of the specific characteristics, quality, and performance of the videos created by MakeItTalk, we recommend referring to the original source and related research work. MakeItTalk's creators have conducted thorough evaluations and validations of their generated content, and their findings provide valuable insights into the capabilities and limitations of the framework. Thus, readers interested in a detailed examination of the video output and the intricacies of the MakeItTalk-generated content are encouraged to explore the relevant sections of the MakeItTalk research literature, where a wealth of pertinent information can be found.

Furthermore, it's crucial to highlight the flexibility inherent in our framework. This adaptability goes beyond being confined to a single talking head generation method, providing users with the freedom to integrate a diverse array of methods into their workflow. While MakeItTalk serves as a fundamental component of our research and has demonstrated its effectiveness, our framework is intentionally designed to accommodate a wide range of state-of-the-art talking head generation approaches.

This versatility translates into expanded opportunities for researchers and practitioners in the field. They are not limited solely to using MakeItTalk but have the flexibility to explore, experiment with, and incorporate alternative methods that align with their specific research goals and requirements. By detaching our framework from reliance on a single method, we empower the research community to harness the full spectrum of innovations and advancements in talking head generation. This, in turn, fosters a more dynamic and diverse landscape of possibilities in multimedia content creation, human-computer interaction, and other related domains.

V. Summary of Results

The complete experimental analysis was performed on a workstation machine equipped with a XEON E5-1650 v4 3.60 GHz processor, 64 GB of RAM, and 2 GEFORCE RTX 2080 graphical processing units each of which has 12 GB of dedicated graphical video memory, memory bandwidth of 616 GB/second, and 4352 cuda cores.

A. *Synthetic Single 2D Child Imaging Facial Data Results*

In the first phase of the experimental analysis, we used distinct boys' and girls' facial data samples which were shortlisted from one of our previous works where we used StyleGAN to tune ChildGAN [4] models for rendering large-scale child synthetic data. Fig. 4 shows some of the child facial samples rendered using the ChildGAN model. Whereas Fig. 5 shows the face localization and 68 facial landmarks detection results on generated synthetic child data using the Dlib library.

B. *Child Speech Synthesizer Results*

The second part includes generating child audio clips using FastPitch architecture. For now, we have written small text sentences which were then used as input feed data for the FastPitch model. Some of these are as below.

1. "It's raining so we will plan some other day".
2. "Overwhelming majority of people in this country know how to discern and differentiate between what they hear and what they read".

3. “London has become one of the most ethnically diverse cities in the world with over 300 languages are spoken in Greater London”.

(a) (b)

Fig. 4. Generated synthetic child facial subjects, LHS: twelve distinct frontal face samples of boys, RHS: twelve distinct frontal face samples of girls.

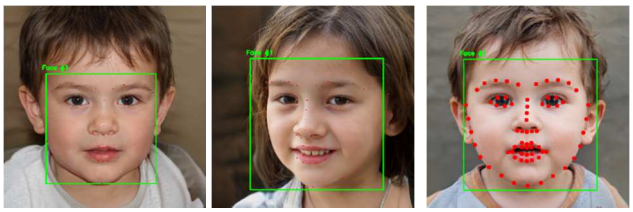


Fig. 5. Face localization and facial landmark detection on ChildGAN data.

Fig. 6 shows the waveform plot of text (“It’s raining so we will plan some other day”) to audio generated file using Fastpitch TTS synthesizer text which is generated in the boy’s audio voice. The audio is plotted with a 2000 maximum number of sampling points.

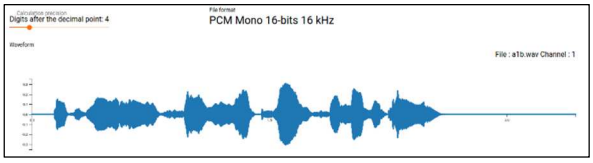


Fig. 6. Audio file plot of the synthesized voice-over of boy generated using Fastpitch TTS.

As mentioned in Section IV-B the generated synthesized voices are further processed by performing a down-sampling operation using the Python librosa library. Fig. 7 demonstrates the graph plot of the original audio WAV file and down-sampled to 16khz audio file.

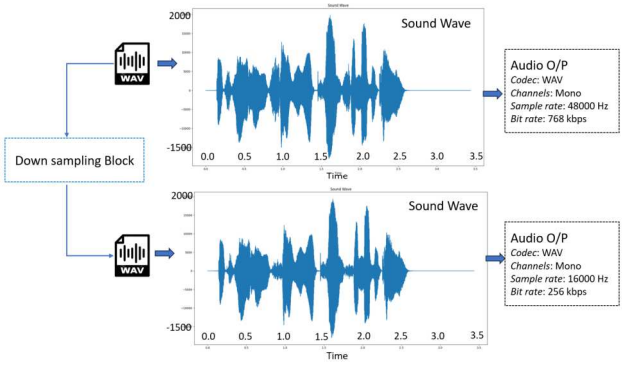


Fig. 7. Downsampled audio WAV file output (“It’s raining so we will plan some other day”) with a time duration of 3.5 seconds, sample rate of 16000 Hz, and bit rate of 256 kbps.

C. Single 2D Facial Image to Talking Child Results

The last phase of the experimental results demonstrates the speaking children videos which are rendered using a single 2D RGB frame and driving audio input. Since this work is in the initial phases now, we have rendered outputs of 20 different child subjects. The synthetic speaking children results along with tuned ChildGAN models used are available on our GitHub repository: https://github.com/MAli-Farooq/Synthetic-3D-Speaking-Children. Fig. 8 shows the selected frame-by-frame results extracted from a rendered speaking child video.

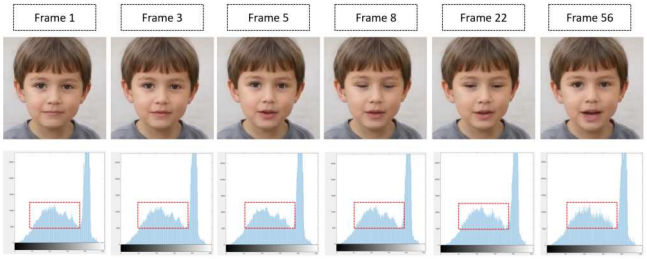


Fig. 8. Five different speaking child video frames of a single subject with their respective histograms.

The second row of Fig. 8 shows the graphical representation of pixel color distribution present in each digital frame which is different from other frames due to continuous lip and eye movement action. Fig. 9 shows the facial landmarks along with varying facial angles of six different facial frames extracted from taking face video results of a similar subject.

D. Subjective Evaluation of Rendered 3D Child Video Data

The quality of rendered data was further evaluated using human subjective evaluation. For this purpose, we have taken the opinions of six participants from our research group by asking them the following questions.

1. Do you agree that the visual quality of the rendered synthetic child video is good?
2. Do you agree that the audio in the video, including speaker similarity, prosody, and audio quality is good?
3. Do you agree the overall video is of natural quality and sharp?

Among this five participants provided the positive response in favor of 1st and 3rd questions whereas four participants agreed with question 2. Thus, on average we got a 75% percent positive response ratio from human evaluation.

VI. Conclusion and Future Work

Our goal in this work has been to demonstrate the potential of synthetic data for replacing “real-world” data in the particular context of a smart-toy platform. To this end, we have combined several advanced data synthesis techniques to provide a working pipeline for speech-driven animated facial training data samples. While this work is still in its early stages the resulting data samples are convincing, capturing realistic facial features such as synchronized lip and jaw movements and eye blinking. Future work will include quantitative evaluation of the uniqueness of the seed facial data samples, improvements in the quality and number of individual speaker embeddings, and improved controllability of the speech generator (e.g. emotional speech embeddings), and further improvements in the quality and controllability of the facial animations. We are also exploring the addition of a diffusion model to the generator for seed facial samples that will allow specific ethnicities, hairstyles, and facial characteristics to be generated.

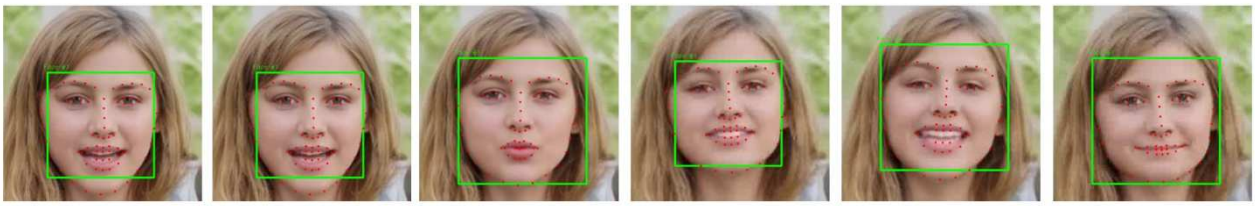


Fig.9: Video frames outputs of girl subject showing varying head pose and lip movements.

ACKNOWLEDGMENTS

This work was supported in part by the DAVID project of the Disruptive Technologies Innovation Fund (managed by the Department of Enterprise, Trade and Employment and administered by Enterprise Ireland), also by the Science Foundation Ireland Centre for Research Training in Digitally Enhanced Reality (www.d-real.ie) under Grant No. 18/CRT/6224, and the ADAPT Centre (Grant 13/RC/2106), and finally by the Irish Research Council Enterprise Partnership Ph.D. Scheme under Grant EPSPG/2020/40.

REFERENCES

[1] G. Costache and P. Corcoran, "DAVID – A Privacy by Design Edge-AI Smart-Toy Platform," Galway, Ireland, Aug. 26, 2023. doi: 10.5281/zenodo.8286232.

[2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[4] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *ACM Trans. Graph. TOG*, vol. 39, no. 6, pp. 1–15, 2020.

[5] M. A. Farooq, W. Yao, G. Costache, and P. Corcoran, "ChildGAN: Large Scale Synthetic Child Facial Data Using Domain Adaptation in StyleGAN," 2023, *arXiv:2307.13746*.

[6] J. Deng, J Guo, N Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogni.*, 2019, pp. 4690-4699.

[7] H. Zhou, S. Hadap, K. Sunkavalli, and D. Jacobs, "Deep Single-Image Portrait Relighting," *in Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7193-7201, doi: 10.1109/ICCV.2019.00729.

[8] Z. Wu, D. Lischinski and E. Shechtman, "StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogni*., 2021, pp. 12858-12867, doi: 10.1109/CVPR46437.2021.01267.

[9] J. J. Burred, E. Ponsot, L. Goupil, M. Liuni and J. J. Aucouturier, "CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition," *PLoS One*, 2019, doi:10.1371/journal.pone.0205943.

[10] A. Łańcucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6588-6592, doi: 10.1109/ICASSP39728.2021.9413889.

[11] J. Shen *et al*., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[12] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran and H. Cucu, "A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," in *IEEE Access*, vol. 10, pp. 47628-47642, 2022, doi: 10.1109/ACCESS.2022.3170836.

[13] L. Wan, Q. Wang, A. Papir and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4879-4883, doi: 10.1109/ICASSP.2018.8462665.

[14] N. Kalchbrenner *et al*., "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.,* 2018, pp. 2415-2424.

[15] A. V. D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *Proc. 9th ISCA Speech Syn. Workshop*, 2016.

[16] H. Zen *et al*., "Libritts: A corpus derived from librispeech for text-to-speech," 2019, *arXiv:1904.02882*.

[17] S. Pradhan, R. Cole, and W. Ward. "MyST Children's Conversational Speech," *LDC2021S05*. Web Download. Philadelphia: Linguistic Data Consortium, 2021.

[18] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 3617-3621, doi: 10.1109/ICASSP.2019.8683143.