

UCLA

UCLA Electronic Theses and Dissertations

Title

ROBUST SPEAKER VERIFICATION UNDER DOMAIN AND DEVELOPMENTAL VARIABILITY:
ADAPTATION AND MODELING IN LOW-RESOURCE CHILD SPEECH

Permalink

<https://escholarship.org/uc/item/93g3m45g>

ISBN

9798190937572

Author

Shetty, Vishwas

Publication Date

2026-09-11

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

ROBUST SPEAKER VERIFICATION UNDER DOMAIN AND DEVELOPMENTAL
VARIABILITY: ADAPTATION AND MODELING IN LOW-RESOURCE CHILD
SPEECH

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Vishwas M. Shetty

ABSTRACT OF THE DISSERTATION

ROBUST SPEAKER VERIFICATION UNDER DOMAIN AND DEVELOPMENTAL VARIABILITY: ADAPTATION AND MODELING IN LOW-RESOURCE CHILD SPEECH

by

Vishwas M. Shetty

Doctor of Philosophy in Electrical and Computer Engineering

University of California, Los Angeles, 2026

Professor Abeer Alwan, Chair

Modern speaker verification systems achieve strong performance on adult speech, but their performance often degrades when applied to children’s speech. This degradation arises from three challenges: children’s speech differs substantially from adult speech, children’s voices continue to change as they grow, and the lack of large child speech databases. Developmental changes affect acoustic and articulatory properties of speech and create age-related variability that can obscure speaker identity. This dissertation addresses the problem of robust speaker verification for children by studying developmental variability in child speech and developing adaptation methods that improve child speaker verification while preserving performance on adult speech.

The first contribution of this dissertation is a longitudinal analysis of acoustic and articulatory development in children. By analyzing vowel formants, subglottal resonances, and ultrasound-based tongue-shape measures from children recorded longitudinally across four years, this dissertation quantifies both systematic developmental trends and child-specific variability. These findings motivate the need for speaker verification systems that are robust to age-related changes in children’s voices. Building on this analysis, the dissertation develops methods for age-robust child speaker verification, including approaches that improve

verification when enrollment and test speech are separated by one or more years. It then addresses the broader problem of efficiently adapting adult speech trained speaker verification models to child speech under limited child speech data.

A central challenge in this progression is catastrophic forgetting, i.e., adaptation to child speech can improve child speaker verification while degrading adult speaker verification performance. This dissertation addresses this challenge first in settings where source-domain data is available, and then in the more restrictive source-free setting where the original adult training data is unavailable during adaptation. Overall, this dissertation provides a progression from quantifying developmental variability in child speech to building robust, efficient, and age-agnostic speaker verification systems that better serve both children and adults.

The dissertation of Vishwas M. Shetty is approved.

Achuta Kadambi

Jonathan Kao

Steven Lulich

Abeer Alwan, Committee Chair

University of California, Los Angeles

2026

To my family, whose unwavering love and support made this journey possible, and especially to my beloved niece and nephews — Anika, Krishiv, Viraansh, and Jayan — who fill my life with joy and inspire me to look forward to the future with hope.

TABLE OF CONTENTS

List of Figures	xiii
List of Tables	xv
Acknowledgments	xvii
Vita	xix
1 Introduction	1
1.1 Speech Technology for Children - Challenges	1
1.2 Developmental Analysis	2
1.2.1 Developmental Variability in Child Speech	2
1.3 Robust Speaker verification systems for Children	3
1.3.1 Age Robust and Efficient Adaptation for Child Speaker Verification	3
1.3.2 Age-Agnostic Speaker Verification	4
1.4 Dissertation Contributions	8
1.5 Dissertation Organization	8
2 Databases and Features	10
2.1 Overview	10
2.2 IU Database - Longitudinal Child Speech Data for Developmental Analysis	10
2.3 Speaker Verification Datasets	11
2.3.1 CSLU Kids' Speech Corpus	11
2.3.2 IU Longitudinal Speaker Verification Evaluation Set	12
2.3.3 My Science Tutor (MyST) Corpus	12

2.3.4	VoxCeleb	13
2.3.5	CNCeleb	13
2.3.6	Dataset Summary	14
2.4	Feature Representations	15
2.4.1	Acoustic Features for Longitudinal Developmental Analysis	15
2.4.2	Articulatory Features for Longitudinal Developmental Analysis	16
2.4.3	Speaker Verification Acoustic Features	16
2.5	Summary	16
3	Experimental Framework, Evaluation Metrics, and Models	17
3.1	Overview	17
3.2	Longitudinal Acoustic and Articulatory Analysis	17
3.2.1	Task Definition	17
3.2.2	Acoustic Analysis Protocol	18
3.2.3	Articulatory Analysis Protocol	18
3.3	Speaker Verification	19
3.3.1	Task Definition	19
3.4	Speaker Verification Models	20
3.4.1	X-vector	20
3.4.2	ECAPA-TDNN	20
3.4.3	ResNet	21
3.4.4	Embedding Extraction and Scoring	22
3.4.5	Evaluation Metric: Equal Error Rate	22
3.5	Data Augmentation and Input Preparation	23
3.6	Training and Adaptation Protocols	23

3.6.1	Training from Scratch	23
3.6.2	Direct Evaluation of pretrained models	23
3.6.3	Fine-Tuning and Target-Domain Adaptation	24
3.7	Evaluation Scenarios	24
3.7.1	Developmental Variability	24
3.7.2	Age-Related Speaker Verification Robustness	24
3.7.3	Adult-to-Child Adaptation	25
3.7.4	Age-Agnostic Verification and Catastrophic Forgetting	25
3.7.5	Source-Free Domain Adaptation	25
3.8	Summary	26
4	Understanding Speech Development in Children	27
4.1	Introduction	27
4.2	Methods	28
4.2.1	Analysis Overview	28
4.2.2	Acoustic Analysis	28
4.2.3	Articulatory Analysis	29
4.2.4	Statistical Comparisons	30
4.3	Results and Discussion	31
4.3.1	Longitudinal Changes in the Acoustic Vowel Space	31
4.3.2	Emergence of Gender-Based Acoustic Differences	33
4.3.3	Development of Subglottal Resonances	34
4.3.4	Relationship Between Tongue Curvature and Vowel Categories	35
4.3.5	Gender-Based and Grade-Based Articulatory Patterns	36
4.3.6	Articulatory Vowel Spaces Across Grades	37

4.3.7	Implications for Child Speaker Verification	39
4.4	Summary	39
5	Enhancing Age-Related Robustness in Children’s Speaker Verification .	41
5.1	Introduction	41
5.2	Methods	42
5.2.1	Age-Related Robustness Evaluation	42
5.2.2	Feature Transform Adapter	43
5.2.3	Synthetic Audio Augmentation	44
5.2.4	Training and Evaluation Setup	44
5.2.5	Models	45
5.3	Results and Discussion	45
5.3.1	Effect of Feature Transform Adapter	45
5.3.2	Effect of Synthetic Audio Augmentation	47
5.3.3	Effect of Training Split and Comparison with Prior C-SV Systems . .	47
5.4	Summary	49
6	Gated Linear Unit Adapter with Iterative Fine-Tuning	50
6.1	Introduction	50
6.2	Methods	51
6.2.1	Adult-to-Child Adaptation Framework	51
6.2.2	Gated Linear Unit Adapter	51
6.2.3	Iterative Fine-Tuning - IFT	52
6.2.4	Experimental Comparison	54
6.3	Results and Discussion	55

6.3.1	Performance Across Speaker Verification Architectures	55
6.3.2	Effect of Target-Domain Training Data Size	57
6.3.3	Discussion	59
6.4	Summary	60
7	Age-Agnostic Speaker Verification	61
7.1	Introduction	61
7.2	Methods	62
7.2.1	Age-Agnostic Speaker Verification (AASV) Framework	62
7.2.2	Domain-Specific Speaker Embedding Networks	63
7.2.3	Age-Domain Representation and Domain Classifier	64
7.2.4	Multi-Space Embedding Integration	64
7.2.5	Experimental Comparison	65
7.3	Results and Discussion	66
7.3.1	Domain Classifier Behavior	66
7.3.2	Adult-Only and Child-Adapted Systems	67
7.3.3	Performance of the Proposed AASV System	68
7.3.4	Ablation Study on Embedding Integration	70
7.3.5	Discussion	70
7.4	Summary	71
8	Cross-Attentive Adapter with Regularized Domain Adaptation for Speaker Verification	73
8.1	Introduction	73
8.2	Methods	74
8.2.1	Source-Free Adaptation Setting	74

8.2.2	Overview of the Proposed Framework	75
8.2.3	Cross-Attentive (CA) Adapter	75
8.2.4	K-Means-Based Elastic Weight Consolidation	78
8.2.5	Moment-Matching Regularization	79
8.2.6	Overall Training Objective	80
8.3	Results and Discussion	80
8.3.1	Evaluation Overview	80
8.3.2	Baseline and Fine-Tuning Behavior	82
8.3.3	Effect of Source-Free Regularization	83
8.3.4	Effect of the Cross-Attentive Adapter	83
8.3.5	Cross-Age Versus Cross-Lingual Adaptation	84
8.4	Summary	85
9	Summary and Future Work	87
9.1	Major Findings and Contributions	88
9.1.1	Quantifying Developmental Variability in Child Speech	88
9.1.2	Improving Age-Related Robustness in Children Speaker Verification	88
9.1.3	Efficient Adult-to-Child Adaptation Under Limited Child Speech Data	89
9.1.4	Building Unified Age-Agnostic Speaker Verification Systems	89
9.1.5	Addressing Catastrophic Forgetting in Source-Free Adaptation	90
9.2	Remaining Challenges	90
9.3	Future Work	91
9.3.1	Larger Longitudinal Acoustic and Articulatory Studies	91
9.3.2	More Realistic Longitudinal Speaker Verification Evaluation	92
9.3.3	Improved Age-Robust and Age-Conditioned Representations	92

9.3.4	Refining Efficient Adaptation Methods	93
9.3.5	Toward Stronger Unified Age-Agnostic Systems	93
9.3.6	Generalizing Source-Free Catastrophic Forgetting Methods	94
9.4	Conclusion	95
References		96

LIST OF FIGURES

4.1	Midsagittal tongue profiles during the production of /u/ in “Zoo” and /æ/ in “Vacuum” by a boy in Grade 1. The tongue contour is highlighted using dotted markings.	30
4.2	Longitudinal changes in the first two formant frequencies from Grade 1 to Grade 4. F1 and F2 form the vertical and horizontal axes, respectively.	31
4.3	Grade-wise changes in the centroids of the vowel spaces for the eight children. The non-monotonic centroid trajectories demonstrate that reductions in formant frequencies with increasing age are not consistent across all children.	32
4.4	Grade-wise average vowel plots highlighting gender-based acoustic differences. .	33
4.5	Grade-wise distributions of the first two subglottal resonances. The mean value is represented by a green triangle marker within each box plot.	34
4.6	Grade-wise vowel plots with F1 and F2 normalized by the first and second subglottal resonances, respectively. Dotted horizontal and vertical lines indicate normalized subglottal resonance frequencies.	35
4.7	Box plots comparing articulatory parameters TCD and TCP. The plots summarize vowel-category, gender-based, and grade-based comparisons. Statistical test results are shown in red in each subplot, and the mean value is represented by a green triangle marker.	36
4.8	Grade-wise TCD versus TCP plots. TCD is plotted along the vertical axis and TCP along the horizontal axis.	38
5.1	Overview of the Feature Transform Adapter used to improve age-related robustness in children speaker verification.	43

6.1	Overview of the G-IFT framework. Panel (a) shows the GLU adapter. Panels (b) and (c) show the iterative fine-tuning procedures for G-IFT-1 and G-IFT-2, respectively.	53
6.2	Absolute reduction in EER (%) by the proposed G-IFT framework (we consider the best result out of G-IFT-1 and G-IFT-2) over conventional fine-tuning across different MyST training splits. Larger reductions are observed in low-resource settings, especially MyST-1 and MyST-2.	58
7.1	Workflow of the Age-Agnostic Speaker Verification (AASV) system. Speaker embeddings E_a and E_c from the A-SV and C-SV subsystems are fused with age-domain information from the domain classifier to generate robust speaker embeddings.	63
7.2	t-SNE visualization of speaker embeddings. Blue and red points denote children’s speech from CSLU and adult speech from VoxCeleb, respectively. (a) Younger children’s group (K00–K06) compared with adult speech. (b) Older children’s group (K07–K10) compared with adult speech.	65
7.3	Performance of the A-SV-Small, C-SV-Small, and proposed AASV-Small systems across year-wise CSLU test sets and the VoxCeleb1 test set.	69
8.1	Overview of the proposed approaches. The CA adapter framework is shown in (a). Hidden representations h_t and h_s have dimension $C \times T$. In the CA Block, “Concat-C” and “Concat-T” refer to concatenation operation on the hidden representations along the C (channel) and T (time) axes, respectively. Estimation of Fisher Information Matrix (FIM) for K-EWC using pseudo-labels / Estimation of pseudo μ_T and Σ_T for MM regularization is shown in (b). D_{Tar} refers to the target domain data.	76

LIST OF TABLES

2.1	Summary of datasets used in this dissertation.	14
5.1	Equal Error Rate (EER) results for systems trained using the CSLU-S split and evaluated on CSLU intra-year trials (Kindergarten - K00 to Grade 10 - K10) and IU intra-year/inter-year trials. Intra-year refers to enrollment and test segments from the same grade, while inter-year refers to enrollment and test segments from different grades.	46
5.2	EER comparisons across combined CSLU, IU intra-year, and IU inter-year evaluation sets. CSLU is obtained by combining K00 to K10 test sets from the CSLU-L evaluation split. IU Intra-Year and IU Inter-Year are obtained by combining the corresponding IU evaluation sets from Table 5.1. [†] indicates training data from CSLU-L, and [*] indicates training data from CSLU-S.	48
6.1	Performance in terms of Equal Error Rate (EER %) across CSLU and MyST-1. <i>Pretrained</i> refers to adult SV systems tested directly on child speech. <i>Baseline</i> refers to systems trained from scratch on child speech. <i>Finetune</i> refers to conventional fine-tuning of adult SV systems using child speech. <i>GLU</i> refers to adapter fine-tuning, and G-IFT-1/G-IFT-2 are the proposed iterative fine-tuning frameworks. Statistical significance is assessed using paired <i>t</i> -tests at $p = 0.05$ and significant improvements over Finetune are marked with “*”.	56
6.2	Equal Error Rate (EER %) across ECAPA-TDNN, ResNet, and x-vector systems on different MyST training splits. The duration of the training splits increases progressively from MyST-1 to MyST-4. The evaluation split is fixed, with no overlap in speakers between training and evaluation sets. Statistical significance is assessed using paired <i>t</i> -tests at $p = 0.05$ and significant improvements over Finetune are marked with “*”.	57

7.1	Accuracy of domain classifiers trained on mixed datasets with varying proportions of adult and child speech. The children test set is a mixture of K00–K10 CSLU test sets, and the adult test set is VoxCeleb1-O.	66
7.2	EER (%) comparison of verification systems. K00–K10 represent increasing child age groups in CSLU. A-SV and C-SV denote adult and children speaker verification systems, respectively. AASV denotes the proposed Age-Agnostic Speaker Verification system. DC denotes the domain classifier, and PreEmbed denotes pretrained speaker embeddings.	68
8.1	Model performance in terms of Equal Error Rate (%EER). Results are reported as Target/Source %EER. For MyST/VoxCeleb-O, the first number is MyST EER and the second number is VoxCeleb-O EER. For CNCeleb/VoxCeleb-O, the first number is CNCeleb EER and the second number is VoxCeleb-O EER. The target training set duration increases from Split-1 to Split-4.	81

ACKNOWLEDGMENTS

This dissertation would not have been possible without the guidance, encouragement, and support of many individuals who have shaped my academic and personal journey.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Abeer Alwan. Her invaluable guidance, unwavering support, and commitment to excellence have been instrumental throughout my doctoral journey. Her mentorship has profoundly influenced the way I approach research, teaching, and academic scholarship, and I feel truly fortunate to have had the opportunity to learn from her.

I am sincerely grateful to Professor Steven Lulich, with whom I had the privilege of collaborating on multiple projects. His expertise, insights, and generosity as a collaborator contributed greatly to my research and intellectual growth. I would also like to express my heartfelt appreciation to Professor Jody Kreiman. Working with her on the psychoacoustic voice-space analysis was an immensely rewarding experience. Her perspective and thoughtful guidance enriched both research and my understanding of voice perception.

I extend my sincere thanks to my committee members, Professor Achuta Kadambi and Professor Jonathan Kao, for their time, thoughtful feedback, and valuable suggestions. Their comments and perspectives helped strengthen this dissertation and encouraged me to consider my research from broader viewpoints.

I am also grateful to Professor Kambiz Shoarinejad, Professor Christina Fragouli, and Professor Ameesh Pandya, for giving me the opportunity to serve as a teaching assistant for their courses. I learned a great deal from each of their distinctive teaching approaches, and those experiences played an important role in shaping my own philosophy and practice as an instructor.

I would like to thank the members of the Electrical and Computer Engineering Department staff, particularly Deona Columbia, Mark Banderas, Tatyana Vigil, Jose Cano, and Tony Velasquez. Their assistance, patience, and kindness helped me navigate the many administrative and practical aspects of graduate school.

I am grateful to my labmates—Amber, Vijay, Jinhan, AJ, Ruchao, Balaji, Eray, Zilai, Kaiyuan, and Mohan—for their companionship, encouragement, stimulating conversations, and all the experiences we shared along the way. I owe special thanks to Jiusi, Anthony, and Injun with whom I had the privilege of collaborating on research and publishing papers. I am particularly grateful for the unique journey I shared with Anthony and Injun who were students in a class for which I served as a teaching assistant and later became research collaborators. Seeing our association grow from the classroom into a productive research collaboration has been especially meaningful to me.

I would also like to thank all the students for whom I served as a teaching assistant. Teaching and interacting with them was a deeply meaningful part of my graduate-school experience. Their questions, curiosity, and enthusiasm continually challenged me to become a clearer communicator and a more thoughtful educator.

Finally, and most importantly, I would like to thank my family. My Amma and Pappa—Srimathi Shetty and Manjunath Shetty, have been my lifelong emotional anchors. Their unconditional love, sacrifices, encouragement, and faith in me have sustained me through every challenge. I consider myself immensely blessed to have them as my parents, and no words can adequately express the depth of my gratitude to them.

To my brother, Shreyas Shetty: you are the reason I have been able to accomplish so much of what I have today. You have always been my role model, my inspiration, and one of my greatest sources of strength. I am equally grateful to my sister-in-law, Namisha Naik. I feel incredibly fortunate to have her as part of my life and my family.

I cannot imagine who or where I would be without all of you in my life. To everyone who stood by my side, believed in me, and supported me throughout this journey, I offer my deepest and most heartfelt thanks. I could not have done any of this without you.

VITA

2021 – 2026	Ph.D. student in Electrical and Computer Engineering, University of California, Los Angeles.
2018 – 2021	M.S. in Electrical Engineering, IIT Madras, India.
2012 – 2016	B.E. in Electronics and Communication Engineering, BMS College of Engineering, Bangalore, India.

PUBLICATIONS

Shetty, V. M., Lulich, S. M., Alwan, A. “Developmental Articulatory and Acoustic Features for Six to Ten Year Old Children.” *In Proc. Interspeech 2023 (pp. 4598-4602).*

Johnson, A., **Shetty, V. M.**, Ostendorf, M., Alwan, A. “Leveraging multiple sources in automatic African American English dialect detection for adults and children.” *In 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096614.*

Shetty, V. M., Zheng, J., Lulich, S. M., Alwan, A. “Enhancing Age-Related Robustness in Children Speaker Verification.” *In 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, doi: 10.1109/ICASSP49660.2025.10888718*

Shetty, V. M., Zheng, J., Alwan, A. “G-IFT: A Gated Linear Unit adapter with Iterative Fine-Tuning for Low-Resource Children’s Speaker Verification.” *In Proc. Workshop on Child Computer Interaction - WOCCI 2025, 1-5, doi: 10.21437/WOCCI.2025-1*

Zheng, J., **Shetty, V. M.**, Shankar, N.B., Alwan, A. “An Age-Agnostic System for Robust

Speaker Verification.” *In Proc. Workshop on Child Computer Interaction - WOCCI 2025*, 41-45, doi: 10.21437/WOCCI.2025-9

Shetty, V. M., Wong, A, Alwan, A. “Cross-Attentive Adapter with Regularized Domain Adaptation for Speaker Verification”, *In 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 18877-18881, doi: 10.1109/ICASSP55912.2026.11463890

CHAPTER 1

Introduction

1.1 Speech Technology for Children - Challenges

Speech-based technologies simplify human interaction with digital systems. Applications such as voice assistants, online learning platforms, personalized educational tools, and security systems rely on speech processing models that can recognize, characterize, or verify users from their voices. Speaker verification (SV), in particular, aims to determine whether two speech samples belong to the same speaker. Modern SV systems have achieved strong performance on large-scale adult speech benchmarks, driven by neural speaker embedding models such as X-vectors [1], ECAPA-TDNN [2], ResNet [3], and related architectures [4, 5]. However, this success has not transferred equally well to children’s speech.

Children’s speech presents a distinct and challenging domain for speech technology. Unlike adult speech, which is relatively stable over time, children’s speech changes substantially as they grow. These changes are driven by developmental variation in vocal tract anatomy, speech motor control, articulation, phonological development, and acoustic realization [6, 7, 8, 9]. As a result, systems trained primarily on adult speech often perform poorly when applied to children. This mismatch is especially problematic for SV, where the goal is to model speaker-specific characteristics while remaining robust to age-related variability. In child speaker verification (C-SV), age-related variability can directly obscure speaker identity, because the same child’s voice may change considerably over the developmental years. This mismatch is compounded by the limited availability of child speech datasets, which makes it difficult to train large neural SV systems directly on child speech.

The central theme of this dissertation is the development of robust speaker verification

systems for children. The dissertation begins by analyzing longitudinal acoustic and articulatory changes in children’s speech. This analysis motivates the core technical problem: children’s voices are not only different from adults’ voices, but also dynamically changing during development. The dissertation then investigates methods to improve C-SV systems under this variability, beginning with age-related robustness within children’s speech, moving to adult-to-child domain adaptation, and finally addressing catastrophic forgetting when adapting adult SV systems to child speech.

1.2 Developmental Analysis

1.2.1 Developmental Variability in Child Speech

Elementary school years are a critical period in a child’s speech and language development. During this period, children’s speech becomes complex, and increasingly adult-like. At the same time, substantial individual variation exists in both acoustic and articulatory behavior. Prior work has examined children’s vowel acoustics, temporal characteristics, and spectral properties, often using cross-sectional designs [6, 7, 8, 9, 10, 11]. However, cross-sectional studies alone cannot fully reveal individual developmental trajectories, because they compare different children at different ages rather than observing how the same child changes over time. Longitudinal analysis is therefore important for understanding child speech development.

Vowel production provides a useful lens for studying these developmental changes. The first two formant frequencies, F1 and F2, are closely related to vowel height and frontness, respectively, and are widely used to characterize vowel spaces. In addition to formants, subglottal resonances (SGRs) provide important information about the interaction between speech acoustics and vocal tract development. Prior work has shown that SGRs can divide the vowel space into high, low, front, and back vowel regions, consistent with quantal theory [12, 13, 14, 15]. However, relatively little longitudinal evidence exists on how these relationships develop during the elementary school years.

Articulatory analysis provides an additional perspective on developmental variability. Acoustic measurements alone do not uniquely determine the articulatory configuration used to produce a sound. Children may achieve similar acoustic targets using different tongue shapes or articulatory strategies. Ultrasound tongue imaging makes it possible to examine midsagittal tongue contours during speech production. This allows developmental analysis to move beyond acoustics and examine how articulatory strategies evolve over time.

The longitudinal analysis in this dissertation uses acoustic and articulatory measurements from children recorded annually from Grade 1 to Grade 4. This analysis shows that acoustic and articulatory development is not uniform across children. The observations from this study motivates the need for speech technologies that explicitly account for developmental variability rather than treating children’s speech as a fixed or homogeneous domain.

1.3 Robust Speaker verification systems for Children

1.3.1 Age Robust and Efficient Adaptation for Child Speaker Verification

In this dissertation we investigate adult-to-child adaptation from two perspectives. First, we study age-related robustness within child speech by developing methods to improve inter-year C-SV performance. Second, we study general low-resource adaptation from adult SV to child SV by using adapter-based fine-tuning strategies that improve knowledge transfer from adult-trained models to child speech. These stages reflect a progression from understanding the variability in child speech to designing adaptation methods that improve C-SV performance under realistic data constraints.

Several strategies have been explored to address the data scarcity problem in C-SV. Feature level adaptation and data augmentation methods attempt to increase training diversity by perturbing adult or child speech. These include speed perturbation, vocal tract length perturbation, noise and reverberation augmentation, SpecAugment-style masking, and voice conversion or synthesis-based approaches that make adult speech more child-like [16, 17, 18, 19]. Such approaches can improve robustness by exposing the model to a wider

range of acoustic conditions. However, augmentation alone does not fully address the mismatch between adult and child speech, especially when the goal is to preserve speaker identity while reducing sensitivity to age-related variation.

Model-level adaptation provides a complementary direction. Rather than training a new model from scratch, adaptation methods modify or extend a pretrained adult SV model using child speech. Fine-tuning, adapter-based learning, domain adaptation losses, and related parameter-efficient methods have been used in speaker verification and speech processing tasks [20, 21, 22, 23, 24, 25, 26, 27]. Adapter-based approaches are particularly attractive in low-resource settings because they can introduce a small number of trainable parameters while preserving much of the pretrained model. This is useful when target-domain child speech is limited and full-model fine-tuning may be unstable or inefficient.

This dissertation proposes novel adapter architectures that ensure better adaptation from adult-to-child domains and also address age-related robustness within C-SV. An efficient fine-tuning technique to deal with data-scarce scenarios for the child speech domain is also proposed.

1.3.2 Age-Agnostic Speaker Verification

Having focused on robust adaptation from adult-to-child adaptation, we next shift our focus towards building age-agnostic SV systems. Age-related robustness and age-agnostic speaker verification are related but distinct goals. Age-related robustness within C-SV focuses on verifying that the speech signal came from a certain child speaker across developmental changes. The enrollment and test utterances may come from the same child but at different grades or ages. In this setting, the model must learn representations that are stable enough to preserve speaker identity while being insensitive to age-related acoustic shifts.

Age-agnostic speaker verification extends this idea further by aiming for strong performance across both children and adults. A model adapted only to children may improve C-SV performance but degrade adult SV performance. Conversely, a model trained only on adult speech may perform well on adult benchmarks but poorly on child speech. A truly

inclusive SV system should therefore support both adults and children without requiring separate systems or sacrificing one domain for the other.

This objective is difficult because adult and child speech differ substantially, and the model must represent speaker identity across domains with different acoustic distributions. Prior domain adaptation work has addressed various mismatch scenarios in speaker verification, including cross-language, cross-device, cross-distance, and cross-dialect conditions [23, 28, 29, 30]. However, age-based mismatch is especially challenging because the acoustic differences between adults and children are tied to biological and developmental factors. Speaker embeddings themselves can contain age-related information [31], which means that age variation may be entangled with speaker identity in the learned representation.

In this dissertation we address this issue by considering systems that combine adult-specific and child-specific speaker representations. The goal is not simply to optimize performance for children, but to construct representations that remain effective across age groups. This dissertation proposes a domain classifier that dynamically combines the representations from adult and child SV models to generate a hybrid embedding that works well on both adult and child domains, hence resulting in an age-agnostic SV system. This provides a bridge between C-SV and broader inclusive speech technology, where systems should work reliably for speakers across age groups.

1.3.2.1 Addressing Catastrophic Forgetting in Domain Adaptation

A major challenge in domain adaptation is catastrophic forgetting. Catastrophic forgetting occurs when a model adapted to a new task or domain, suffers from performance degradation on the original task or domain [32, 33]. In the context of this dissertation, an adult-trained SV model adapted to child speech may improve child SV performance while degrading adult SV performance. This creates a stability-plasticity problem: the model must be plastic enough to learn the child speech domain, but stable enough to retain source-domain adult speaker knowledge.

Catastrophic forgetting has been studied in continual learning, where methods are often

grouped into architectural, rehearsal-based, and regularization-based approaches. Architectural methods modify or expand the network to reduce interference across tasks [34, 35, 36, 37, 38]. Rehearsal-based methods retain or replay data from previous tasks during adaptation [39]. Regularization-based methods constrain important parameters or outputs so that the model does not drift too far from the source task [40, 41, 42, 43, 44]. These strategies provide useful tools for balancing adaptation and retention. In the context of this dissertation, if adapting the model to children causes a large degradation on adult speech, the adapted system becomes less useful as a general-purpose verifier.

This dissertation, considers a challenging adaptation scenario, where source data is unavailable during adaptation. Source-free domain adaptation addresses the case where a pretrained source-domain model is available, but the original source-domain data is not. This setting is realistic for speech applications because large-scale training datasets may be proprietary, restricted, or difficult to store and redistribute. In such cases, the model must adapt to the target domain using only the pretrained model and target-domain data. This constraint makes catastrophic forgetting more difficult to address.

Although effective, standard rehearsal-based methods [39] are incompatible with the source-free setting considered in this dissertation because they require continued access to source data. Generative-replay methods avoid directly storing source examples by training a generative model to synthesize samples representative of previously learned tasks [45]. However, they require an additional generator to be trained on the source domain and retained alongside the source model. Consequently, generative replay does not satisfy the stricter setting considered in this dissertation.

Knowledge distillation (KD) [43] transfers information from a fixed teacher model to a student model by encouraging the student to reproduce the teacher’s softened output distribution. However, the effectiveness of this transfer depends on whether the teacher produces informative outputs for the data used during distillation. In the present setting, an adult-trained speaker model is evaluated using child speech, for which a substantial domain mismatch exists. Consequently, the teacher outputs on the child data may be poorly

calibrated or may exhibit limited variation across different child speakers.

Learning without Forgetting (LwF) [46] applies knowledge distillation to continual learning without requiring access to source-domain examples. Before adaptation, the frozen source model is applied to the target-domain samples, and its predictions for the original task are recorded as soft targets. During adaptation, the model is jointly optimized using the new-task objective and a distillation loss that encourages its predictions for the original task, evaluated on the target-domain samples, to remain close to those produced by the frozen source model. Thus, LwF preserves the source model’s behavior on the target samples while allowing the shared representation to be updated for the new task. Like in KD, severe domain mismatch between adult and child speech domains can cause these predictions to be poorly calibrated, overly concentrated, or insufficiently diverse across target speakers. In such cases, the LwF objective may encourage the adapted model to preserve an unreliable or weakly discriminative teacher response.

Parameter-regularization methods like Synaptic Intelligence (SI) [41] estimates parameter importance by accumulating each parameter’s contribution to changes in the training loss throughout source-task optimization. Although SI is computationally efficient, it requires access to the source-training trajectory or importance statistics saved during that training. These quantities are unavailable when adaptation begins with only the final source-model checkpoint.

In this dissertation, we address source-free adaptation for speaker verification under both cross-age (Adult $\rightarrow$ Child) and cross-lingual (English $\rightarrow$ Chinese) domain shifts. The key idea is to exploit pretrained-model embeddings extracted from target-domain inputs as a proxy for source-domain knowledge. These embeddings retain information from the source-trained model even though the original source data is unavailable. This enables the design of adaptation strategies that constrain the target-domain model while reducing source-domain degradation. By addressing catastrophic forgetting under a zero-source-data constraint, the dissertation moves toward more practical and deployable adult-to-child speaker verification frameworks.

1.4 Dissertation Contributions

The main contributions of this dissertation are as follows.

- Presents a longitudinal analysis of developmental acoustic and articulatory variability in children’s speech.
- Introduces evaluation and adaptation methods for age-related robustness in child speaker verification. Examines inter-year C-SV, where enrollment and test utterances are separated by one or more years, and studies methods that improve robustness to developmental changes in the same child’s voice.
- Develops adapter-based adult-to-child adaptation methods for low-resource C-SV. By leveraging adult-trained speaker verification models and adapting them using limited child speech, this dissertation studies how pretrained adult SV knowledge can be transferred efficiently to the child speech domain.
- Investigates age-agnostic speaker verification, where the goal is to maintain strong performance across both adult and child speech.
- Addresses catastrophic forgetting in adaptation methods assuming no access to source domain data, i.e., adult speech. This dissertation proposes methods that reduce source-domain degradation in a source-domain data free adaptation framework.

1.5 Dissertation Organization

The remainder of this dissertation is organized as follows. Chapter 2 describes the datasets and feature representations used across this dissertation. Chapter 3 presents the experimental framework, including speaker verification models, evaluation metrics and training protocols. Chapter 4 presents the longitudinal acoustic and articulatory analysis of children’s speech. Chapter 5 investigates methods for improving age-related robustness in child speaker

verification. Chapter 6 introduces adapter-based iterative fine-tuning for low-resource adult-to-child speaker verification adaptation. Chapter 7 presents an age-agnostic speaker verification framework designed to improve performance across both adult and child speech. Chapter 8 addresses catastrophic forgetting under source-free domain adaptation. Finally, Chapter 9 summarizes the dissertation and discusses future research directions.

CHAPTER 2

Databases and Features

2.1 Overview

This chapter, describes the databases and features used in this dissertation. The aim of this dissertation was to study the effect of age-related acoustic variability on child speech processing, with a focus on speaker verification. The experimental work uses two broad categories of data. The first category consists of longitudinal child speech production data, used to characterize developmental changes in acoustic and articulatory features. The second category consists of speaker verification corpora, used to study age-related robustness, adult-to-child domain adaptation, age-agnostic speaker verification, catastrophic forgetting, and source-free domain adaptation.

The organization of this chapter is as follows: Section 2.2 describes the longitudinal child speech data used to analyze developmental acoustic and articulatory variability. Section 2.3 introduces the datasets used across the remaining chapters on speaker verification. Section 2.4 summarizes the feature representations used in all experiments.

2.2 IU Database - Longitudinal Child Speech Data for Developmental Analysis

The first component of this dissertation examines how child speech changes during the elementary school years. This analysis uses a longitudinal child speech production dataset referred to as IU database [47, 48] collected by our project collaborators at Indiana University. Data were collected from eight typically developing children: four female and four

male native speakers of Midwestern American English. The children were recorded once per year from Grade 1 through Grade 4, covering approximately ages 6 to 10. All children had Goldman-Fristoe Test of Articulation, Third Edition (GFTA-3), standardized scores within the normal range, and age and gender information were recorded for each participant.

The data were collected using synchronized microphone, neck accelerometer, and ultrasound recordings. The microphone signal was used for acoustic vowel analysis, the neck accelerometer signal was used for estimating subglottal resonances, and the ultrasound images were used for midsagittal tongue contour analysis. More details of this database will be provided in 4.

2.3 Speaker Verification Datasets

The speech datasets used in our speaker verification experiments are discussed in the following subsections.

2.3.1 CSLU Kids’ Speech Corpus

The CSLU Kids’ Speech Corpus [49], also referred to as the OGI Kids’ Speech Corpus, contains speech from approximately 1,100 children, spanning kindergarten through Grade 10, with roughly 100 speakers per age group. The corpus includes both spontaneous and scripted speech. All the experiments in this dissertation, use only the scripted speech portion of the dataset.

Two CSLU train-evaluation splits are used in our experiments. One split uses a relatively large training set, with 866 speakers and approximately 24 hours of speech, and an evaluation set spanning 255 speakers covering 44,000 speaker verification trials. This split is referred to as *CSLU-S*, where the suffix “*S*” indicates presence of smaller number of speaker in the evaluation set. The second train-evaluation split follows a low-resource protocol, with 120 training speakers, approximately spanning 2.65 hours of speech, and an evaluation set containing 993 speakers and 190,972 verification trials. This split is referred to as CSLU-L

[19] in this dissertation, where the suffix “*L*” indicates larger number of speakers in the evaluation set. In both train-eval splits, train and evaluation speakers are non-overlapping. Age labels in the CSLU corpus, supports evaluation of age-group-specific speaker verification.

2.3.2 IU Longitudinal Speaker Verification Evaluation Set

The longitudinal child speech production dataset [47, 48] described in Section 2.2 is also used to construct a speaker verification evaluation set. This evaluation set is referred to here as the IU longitudinal evaluation set. Verification trials are then constructed to evaluate both intra-year and inter-year speaker verification across Grade 1 (G1), Grade 2 (G2), Grade 3 (G3), and Grade 4 (G4).

The intra-year sets compare enrollment and test segments from the same grade, producing conditions such as G1–G1, G2–G2, G3–G3, and G4–G4. The inter-year sets compare enrollment speech from one grade with test speech from later grades, producing conditions such as G1–G2, G1–G3, G1–G4, G2–G3, G2–G4, and G3–G4. Each evaluation set contains 2,000 verification trials, consisting of 1,000 positive trials (same speaker) and 1,000 negative trials (different speakers). This dataset is especially important for evaluating the effect of age-related voice changes on child speaker verification.

2.3.3 My Science Tutor (MyST) Corpus

The MyST corpus [50] is used as a larger child speech dataset in adult-to-child adaptation and low-resource child speaker verification experiments. The full corpus contains speech totalling 499 hours from 1,372 speakers in Grades 3 to 5 without explicit grade labels unlike CSLU corpus [49]. In the experiments considered in this dissertation, the annotated portion of the corpus is used, totaling approximately 268 hours of speech.

Multiple MyST training subsets are constructed to study the effect of the amount of in-domain child speech available for adaptation. The subsets are referred to as MyST-1, MyST-2, MyST-3, and MyST-4. Each subset contains data from all 1,210 training speakers,

while the amount of speech per speaker increases across the subsets. The approximate durations across MyST-1, MyST-2, MyST-3, and MyST-4 subsets are 2 hours, 8 hours, 85 hours, and 268 hours, respectively. A fixed evaluation set with 91 speakers and 100,000 verification trials is used across these four training conditions. There is no overlap between training and evaluation speakers.

The MyST subsets are useful for evaluating adaptation under different resource conditions. The smaller MyST-1 subset represents an extremely low-resource adaptation scenario, while the larger MyST-4 subset enables a more data-rich child adaptation setting.

2.3.4 VoxCeleb

VoxCeleb [51, 52] is the primary adult speaker verification dataset consisting of large-scale English speech data collected from YouTube and includes substantial variation in speakers, acoustic environments, and recording conditions. In our experiments, we do not train speaker verification models using Voxceleb, instead only models trained on VoxCeleb are used as pretrained models before adaptation to child speech. However, the VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H evaluation sets are used as a standard adult speaker verification evaluation set to quantify catastrophic forgetting in terms of adult-speaker verification performance degradation post adaptation.

2.3.5 CNCeleb

CNCeleb [53] is a large scale adult speaker verification dataset in Chinese. This dataset contains speech from Chinese open-media sources. It is used in the domain adaptation experiments in Chapter 8 as an additional target-domain dataset. Unlike MyST, which is used for cross-age adaptation from adult English to child English, CNCeleb is used for cross-lingual adaptation from English to Chinese adult speech. In the experiments considered in Chapter 8, only CNCeleb training segments of at least three seconds are retained, and shorter files are discarded leading to a training set with 162 hours of speech from 781 speakers.

Four CNCeleb training subsets are constructed, referred to as CNCeleb-1, CNCeleb-2, CNCeleb-3, and CNCeleb-4. These subsets contain all 781 speakers, with increasing amounts of training speech: approximately 1.94 hours, 10.00 hours, 98.55 hours, and 162.20 hours, respectively. Evaluation is performed using the standard CNCeleb evaluation set. CNCeleb is included to test whether adaptation strategies designed for catastrophic forgetting are effective beyond the adult-to-child setting.

2.3.6 Dataset Summary

The details of the major datasets used in this dissertation and their primary role are summarized in Table 2.1.

Table 2.1: Summary of datasets used in this dissertation.

Dataset	Domain	Main statistics	Role in dissertation
Longitudinal IU child production data	Scripted Speech - Child speech production	8 children; 4 female and 4 male; Grade 1-4; yearly recordings	Developmental acoustic and articulatory analysis
IU longitudinal SV set	Scripted Speech - Child speaker verification	Grade 1-4 intra-year and inter-year trials;	Evaluation of age-related robustness in child speaker verification
CSLU	Scripted Speech - Child speaker verification	Approximately 1,100 children; K00–K10; scripted portion used	Child speaker verification, adult-to-child adaptation, age-group evaluation

Continued on next page

Table 2.1: Summary of datasets used in this dissertation. Continued.

Dataset	Domain	Main statistics	Role in dissertation
MyST	Spontaneous Speech - Child speaker verification	Annotated portion approximately 268 h; training subsets of 2 h, 8 h, 85 h, and 268 h from 1,210 speakers; fixed 91-speaker evaluation set	Low-resource and high-resource child adaptation experiments
VoxCeleb	Spontaneous Speech - Adult English speaker verification	VoxCeleb1 and VoxCeleb2 pretrained SV model used; evaluations on VoxCeleb1-O/E/H	Adult source-domain pretrained model and catastrophic forgetting evaluation
CNCeleb	Spontaneous Speech - Adult Chinese speaker verification	Four training subsets: 1.94 h, 10.00 h, 98.55 h, and 162.20 h from 781 speakers after duration filtering; evaluated on standard 200 speaker evaluation set	Cross-lingual adaptation

2.4 Feature Representations

2.4.1 Acoustic Features for Longitudinal Developmental Analysis

The longitudinal developmental analysis uses acoustic features that characterize vowel production and vocal tract growth. The first two formant frequencies, F1 and F2, are manually

measured from the microphone recordings using Praat. The first two subglottal resonances, Sg1 and Sg2, are manually measured from the neck accelerometer signal.

2.4.2 Articulatory Features for Longitudinal Developmental Analysis

The articulatory analysis uses ultrasound-derived midsagittal tongue contours. For each relevant ultrasound image, the midsagittal tongue contour is manually extracted. The contour is then summarized using two articulatory measures: Tongue Curvature Degree (TCD) and Tongue Curvature Position (TCP) [54, 55, 56]. TCD measures the degree of tongue curvature, while TCP measures the position of the curvature along the tongue contour. These measures are useful because they are relatively robust to variation in ultrasound probe placement and provide compact descriptors of the tongue shape. In the developmental analysis, TCD is used to examine vowel height-related articulatory differences, while TCP is used to examine vowel frontness-related differences.

2.4.3 Speaker Verification Acoustic Features

For the speaker verification experiments, the common acoustic input representation is an 80-dimensional filterbank feature sequence. The filterbank features are extracted using a 25 ms analysis window and a 10 ms frame shift.

2.5 Summary

In this chapter the datasets and feature representations used throughout the dissertation were introduced. The longitudinal child speech production data provide the empirical basis for studying developmental variability in acoustic and articulatory speech features. The speaker verification datasets support a sequence of experiments on age-related robustness, age-agnostic speaker verification, and adaptation. The next chapter defines the experimental framework, evaluation metrics, and training protocols used.

CHAPTER 3

Experimental Framework, Evaluation Metrics, and Models

3.1 Overview

In this chapter we describe the common experimental framework used in this dissertation. The goal is to define the tasks, model families, evaluation metrics, and training protocols that are shared across experiments.

3.2 Longitudinal Acoustic and Articulatory Analysis

3.2.1 Task Definition

The longitudinal analysis is designed to characterize how children produce vowels across Grades 1 to 4. The analysis focuses on both acoustic and articulatory development. The acoustic analysis examines changes in vowel formants and subglottal resonances, while the articulatory analysis examines changes in tongue shape using ultrasound-derived measures.

This analysis provides the developmental motivation for the speaker verification work in the later chapters: i.e., if a child’s acoustic and articulatory speech characteristics change across age, then a child speaker verification system trained or enrolled at one age may not remain robust when tested at a later age.

3.2.2 Acoustic Analysis Protocol

The acoustic analysis uses the first two formant frequencies, F1 and F2, measured from microphone recordings. Measurements are made for four corner vowels, each appearing in two word contexts, across the four yearly (Grade 1 to 4) recordings. This produces repeated measurements for each child across vowel, word, and grade.

The formant measurements are visualized using F1–F2 vowel plots. These plots are examined at the individual-child level, at the grade level, and at the gender level. Individual-child plots reveal longitudinal trajectories that may differ across children, while grade-averaged and gender-averaged plots reveal broader developmental trends.

The first two subglottal resonances, Sg1 and Sg2, are measured from the neck accelerometer signal. Because subglottal resonances are treated as vowel-independent for a given child and grade, they provide a developmental reference for interpreting vowel space organization. The analysis therefore includes SGR-normalized vowel spaces, where F1 and F2 are normalized with respect to Sg1 and Sg2. These normalized plots are used to examine whether the subglottal resonances separate high, low, front, and back vowel regions during development.

3.2.3 Articulatory Analysis Protocol

The articulatory analysis uses midsagittal tongue contours extracted from ultrasound images. Each contour is summarized using Tongue Curvature Degree (TCD) and Tongue Curvature Position (TCP) [54]. TCD is used as a compact measure of the degree of tongue-body curvature, while TCP is used as a compact measure of where the curvature occurs along the tongue contour.

The analysis compares TCD values for high versus low vowels and TCP values for front versus back vowels. This comparison tests whether the articulatory measures correspond to known vowel production distinctions. TCD and TCP are also compared across gender and grade to examine whether systematic developmental or sex-based patterns are present in the articulatory data.

3.2.3.1 Statistical Comparisons

The longitudinal analysis uses descriptive plots and statistical tests. Box plots are used to summarize the distributions of TCD, TCP, Sg1, and Sg2 across relevant groupings. Two-sample one-tailed t-tests are used for targeted comparisons such as high versus low vowels for TCD and front versus back vowels for TCP. One-way analysis of variance (ANOVA) is used to test for main effects of grade on articulatory measures.

These statistical tests are used to support developmental interpretation rather than to train predictive models. The purpose of this part of this dissertation is therefore explanatory: it establishes that child speech development involves measurable acoustic and articulatory changes, and that these changes can vary across children.

3.3 Speaker Verification

3.3.1 Task Definition

Speaker verification is the task of deciding whether two speech samples belong to the same speaker. In a typical trial, one sample or set of samples is treated as the enrollment speech, and another sample is treated as the test speech. The system extracts speaker embeddings from the enrollment and test samples, computes a similarity score, and compares that score to a decision threshold.

A target (or positive) trial is a trial in which the enrollment and test samples come from the same speaker. A non-target (or negative) trial is a trial in which the enrollment and test samples come from different speakers. The verification system should assign higher scores to target trials than to non-target trials.

3.4 Speaker Verification Models

The first step in building a speaker verification model is to train a deep neural network on the task of speaker identification (SID). In order to perform speaker verification, fixed dimensional feature vectors are extracted from these models trained on the SID task. The feature vectors extracted are the speaker embeddings discussed in Section 3.3.1 In this dissertation, we report our findings on multiple deep neural models, in order to highlight the applicability of our proposed approaches across network architectures. We use X-vector [1], ECAPA-TDNN [2], and ResNet [3] based models in our experiments.

3.4.1 X-vector

The X-vector system is based on a Time Delay Neural Network (TDNN) architecture [57]. The TDNN layers operate on short-term acoustic (filterbank) features and capture speaker-relevant patterns over local temporal contexts. In the lower TDNN layers, the network captures information from a relatively small number of neighboring frames. As multiple TDNN layers are stacked, the effective temporal context increases, allowing deeper layers to represent speaker characteristics over a wider span of the speech signal. These frame-level representations are then passed to a statistics pooling layer, which converts the variable-length sequence of frame-level features into a fixed-dimensional segment-level representation. The statistics pooling layer computes the mean and standard deviation of the frame-level representations across time, thereby summarizing both the average characteristics and temporal variability of the speaker’s voice within an utterance. This pooled representation is subsequently passed through fully connected layers to produce the X-vector, a compact fixed-dimensional speaker embedding.

3.4.2 ECAPA-TDNN

ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Network) [2] extends the standard TDNN-based X-vector architecture by in-

roducing a more expressive frame-level feature extraction network and a stronger pooling mechanism.

A key component of ECAPA-TDNN is the use of Res2Net-style [58] TDNN blocks, which divide the feature channels into multiple groups and process them hierarchically. This allows the network to capture speaker-relevant information over multiple temporal scales within a single block. As in the X-vector system, stacking multiple TDNN-based layers further increases the effective temporal context. However, ECAPA-TDNN strengthens this process by propagating and aggregating information from multiple intermediate layers, allowing the final representation to retain both lower-level and higher-level acoustic cues.

ECAPA-TDNN also introduces channel attention through squeeze-and-excitation [59] mechanisms. These mechanisms allow the network to dynamically emphasize channels that are more informative for speaker discrimination while suppressing less useful or noisy channels. In addition, in order to generate a fixed speaker embedding, ECAPA-TDNN improves upon standard statistics pooling by using channel-dependent attentive statistics pooling [60], where the importance of each frame can vary across feature channels. This produces a more informative speaker embeddings than simple unweighted mean and standard deviation pooling done in standard X-vector systems.

3.4.3 ResNet

ResNet-based speaker verification systems [3] use convolutional neural networks with residual connections to extract speaker-discriminative representations. Unlike TDNN-based systems, which model temporal context through one-dimensional temporal operations, ResNet architectures typically operate on two-dimensional time-frequency representations and learn local patterns across both time and frequency.

The key component of ResNet is the residual block, where the input to a block is added to the output of a sequence of convolutional layers. This skip connection allows the network to learn residual mappings rather than entirely new transformations, making it easier to train deeper networks and reducing problems such as vanishing gradients. As the signal

passes through successive residual blocks, the network learns increasingly abstract acoustic representations that capture speaker-relevant spectral and temporal patterns.

For speaker verification, the frame-level or feature-map representations produced by the convolutional layers are converted into a fixed-dimensional speaker embedding using a pooling or aggregation layer.

3.4.4 Embedding Extraction and Scoring

As discussed in the earlier subsections, the speaker verification systems map acoustic filterbank features to fixed-dimensional speaker embeddings. These embeddings are then compared using cosine similarity in the verification backend. Cosine scoring is appropriate because the speaker embedding is intended to encode speaker identity information in a vector space where same-speaker pairs should be close and different-speaker pairs should be separated.

In the X-vector, ECAPA-TDNN, and ResNet based experiments, the speaker embedding dimensions are typically 512, 192, and 256 respectively. The same general cosine similarity based scoring framework is used across the child speaker verification, adult-to-child adaptation, age-agnostic verification, and catastrophic forgetting experiments.

3.4.5 Evaluation Metric: Equal Error Rate

The main speaker verification metric used throughout this dissertation is Equal Error Rate (EER). The EER is the operating point at which the false acceptance rate (FAR) equals the false rejection rate (FRR). Lower EER indicates better speaker verification performance.

For a set of verification scores, the FAR measures the proportion of non-target trials incorrectly accepted as target trials, while the FRR measures the proportion of target trials incorrectly rejected. By sweeping the decision threshold, one obtains a trade-off between these two error rates. The Equal Error Rate (EER) is the error rate at the threshold where FAR and FRR are equal. Thus, EER provides a single summary of verification performance

without requiring a predefined application-specific decision threshold.

3.5 Data Augmentation and Input Preparation

Data augmentation is used to improve robustness in all our speaker verification experiments. The common augmentation methods include additive noise, room impulse response reverberation, frequency masking or frequency dropping, and time masking or time dropping [61, 62, 63, 64]. In the experiments, noise augmentation uses signal-to-noise ratios between 0 and 15 dB, frequency dropping removes one to three frequency bands, and time dropping removes one to five chunks of 1000 to 2000 samples per audio sample.

Input segment duration constraints were set based on available sample durations present in a given dataset. For example, in child speaker verification experiments (MyST and CSLU) we only use two-second speech segments, and in the case of CNCeleb speaker verification experiments we retain only training segments of at least three seconds.

3.6 Training and Adaptation Protocols

3.6.1 Training from Scratch

Our experiments include systems trained from scratch on the target-domain training data. These systems provide a baseline for evaluating whether transfer from an adult pretrained model is useful. We evaluate training from scratch under varied train data-availability scenarios. For example low-resource setting using CSLU-L (2.65 hours), MyST-1 (2 hours), or CNCeleb-1 (2 hours). Similarly, relatively high-resource scenario can be simulated using MyST-4 (268 hours) and CNCeleb-4 (162 hours).

3.6.2 Direct Evaluation of pretrained models

Adult pretrained speaker verification models are used as source-domain systems. These models trained on VoxCeleb, are either evaluated directly on target domain or adapted to

target domain. Direct evaluation of adult-trained systems on child data measures the severity of adult-to-child mismatch (or language mismatch in case of adult-English to adult-Chinese). Poor performance in this setting motivates the need for adaptation.

3.6.3 Fine-Tuning and Target-Domain Adaptation

Fine-tuning adapts an adult pretrained speaker verification model using target-domain data, such as CSLU or MyST for child speaker verification and CNCeleb in case of Chinese speaker verification. In the general framework, the pretrained speaker embedding model is initialized from the VoxCeleb adult speaker verification model, and the system is optimized using target-domain speaker labels. In this dissertation, we propose various efficient finetuning strategies, which will be discussed in detail in subsequent chapters.

3.7 Evaluation Scenarios

3.7.1 Developmental Variability

The first evaluation scenario examines developmental variability in child speech. It uses longitudinal acoustic and articulatory measurements to compare vowels, grades, gender groups, and individual child trajectories. The goal is to show that child speech development is not only a group-level phenomenon but also involves child-specific patterns. This will be the focus of our discussion in Chapter 4.

3.7.2 Age-Related Speaker Verification Robustness

The second evaluation scenario tests child speaker verification under intra-year and inter-year conditions. Intra-year trials measure performance when enrollment and test speech are drawn from the same grade (i.e., age group). Inter-year trials measure performance when enrollment and test speech are separated by one or more years. The difference between these conditions quantifies the impact of age-related voice changes on speaker verification systems.

This will be the focus of our discussion in Chapter 5.

3.7.3 Adult-to-Child Adaptation

The third evaluation scenario evaluates adaptation strategies for adaptation from adult speaker verification models to child speaker verification models. This setting compares adult pretrained systems, child-domain systems trained from scratch, and adapted systems. The key evaluation criterion is whether adaptation improves EER on child speaker verification tasks. This will be the focus of our discussion in Chapter 6.

3.7.4 Age-Agnostic Verification and Catastrophic Forgetting

The fourth evaluation scenario considers both child and adult speaker verification performance. A system is age-agnostic if it performs competitively on both child and adult speech. This setting requires reporting target-domain performance on child speech and source-domain performance on adult speech. The gap between adapted model performance and source model performance on source evaluation set, i.e., adult evaluation data is central to the catastrophic forgetting analysis. This will be the focus of our discussion in Chapter 7.

3.7.5 Source-Free Domain Adaptation

The final evaluation scenario considers adaptation assuming no access to source-domain data. The model is adapted using only target-domain data, but it is still evaluated on both the target domain and the original source-domain evaluation set. Similar to evaluation scenario 3.7.4, the model is evaluated based on performance on both target and source evaluation sets. The emphasis here is on the trade-off between target-domain adaptation and source-domain performance retention. This will be the focus of our discussion in Chapter 8.

3.8 Summary

This chapter defined the common experimental framework used in this dissertation. The longitudinal acoustic and articulatory analysis establishes the developmental variability problem. The speaker verification framework provides a shared evaluation structure for the remaining chapters. Later chapters build on this framework to introduce specific methods for age-related robustness, adult-to-child adaptation, age-agnostic speaker verification, catastrophic forgetting, and source-free adaptation.

CHAPTER 4

Understanding Speech Development in Children

4.1 Introduction

The goal of this chapter is to quantify longitudinal acoustic and articulatory variability in children's speech. The analysis focuses on four corner vowels, /æ/, /i/, /u/, and /ɑ/, observed from Grade 1 (approximately age 6) to Grade 4 (approximately age 10). These vowels provide a compact way to study changes in vowel height and frontness, and therefore provide a useful window into developmental changes in the vowel space. The acoustic analysis examines the first two formant frequencies - F1 and F2, and the first two subglottal resonances - Sg1, and Sg2. The articulatory analysis examines tongue-shape measures derived from midsagittal ultrasound tongue contours.

The main questions addressed in this chapter are:

- How do the acoustic vowel spaces of children change from Grade 1 to Grade 4?
- Do gender-based acoustic differences emerge during this period?
- How do subglottal resonances change across grades, and how do they relate to the organization of the vowel space?
- Are tongue curvature measures systematically related to vowel height and vowel frontness?
- Do the articulatory patterns show the same developmental trends as the acoustic patterns?

The measured acoustic and articulatory features in this chapter, characterize developmental variability in the respective aspects of child speech. The findings provide the empirical motivation for the age-robust and adult-to-child speaker verification methods developed in later chapters.

4.2 Methods

4.2.1 Analysis Overview

The database and feature representations used in this chapter were introduced in Chapter 2. Briefly, the analysis uses longitudinal recordings from eight typically developing children, consisting of four female and four male native speakers of Midwestern American English. The children were recorded once per year from Grade 1 to Grade 4. The data available include - speech transcriptions, microphone recordings, neck accelerometer recordings, and ultrasound tongue images.

For each child and grade, the analysis considers four corner vowels: /æ/, /i/, /u/, and /ɑ/. Each vowel is recorded in two word contexts - “Apple”/“Vacuum” (/æ/), “Teeth”/“Zebra” (/i/), “Shoe”/“Zoo” (/u/) and, “Watch”/“Frog” (/ɑ/). For the acoustic analysis, formant frequencies were measured from the microphone signal, and subglottal resonances were measured from the accelerometer signal. For the articulatory analysis, midsagittal tongue contours were extracted from the mid-sagittal view of the ultrasound images aligned with the corresponding speech signal.

4.2.2 Acoustic Analysis

The first two formant frequencies, F1 and F2, were used to characterize the acoustic vowel space. For each vowel in each grade, the two word contexts provided two measurements of F1 and F2. To make the measurements, for each target vowel, a temporally stable segment was selected to minimize the influence of vowel onset, offset, and coarticulatory transitions. Formant frequencies were initially identified from the broadband spectrogram and subse-

quently cross-checked against the spectral envelope obtained from the same stable vowel segment. This combined spectrographic and spectral inspection was particularly important for productions with a high f_0 , where the wide spacing between harmonics can make formant estimation less reliable. In such cases, the spectral representation was used to distinguish individual harmonics from broader vocal-tract resonances and to verify the formant values observed in the spectrogram. These measurements were averaged when constructing grade-wise vowel plots for individual children and gender-based group averages.

The first two subglottal resonances, Sg1 and Sg2, were analyzed as speaker-level measures for each recording year. Since subglottal resonances are treated as vowel-independent properties for a given child and grade [65], they provide reference values for interpreting the organization of the acoustic vowel space. To examine this relationship, F1 and F2 were normalized with respect to Sg1 and Sg2, respectively, and the normalized vowel spaces were visualized across grades.

4.2.3 Articulatory Analysis

The articulatory analysis uses Tongue Curvature Degree (TCD) and Tongue Curvature Position (TCP), which summarize the shape and position of the midsagittal tongue contour. The use of TCD and TCP follows prior work showing that such measures are relatively robust to ultrasound probe-placement variability and can characterize tongue-shape differences in speech production [54, 55, 56].

Figure 4.1 illustrates the midsagittal tongue profiles for /u/ and /æ/ and the geometric basis for TCD and TCP. Let X and Y denote the tongue tip and tongue root, respectively. Let Z be the point on the tongue contour with the maximum perpendicular distance from the line XY, and let O be the projection of Z onto XY. TCD and TCP are computed as

$$TCD = \frac{\text{len}(ZO)}{\text{len}(XO)} = \frac{r}{p}, \quad TCP = \frac{\text{len}(YO)}{\text{len}(XO)} = \frac{q}{p}. \quad (4.1)$$

Here, $\text{len}(\cdot)$ refers to length, TCD represents the degree of tongue curvature, while TCP represents the position of the curvature along the tongue contour.

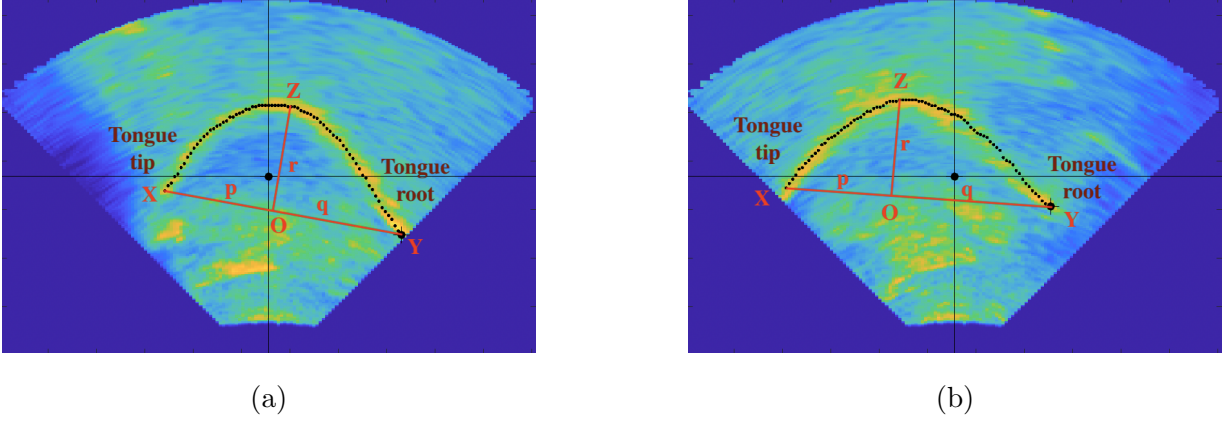
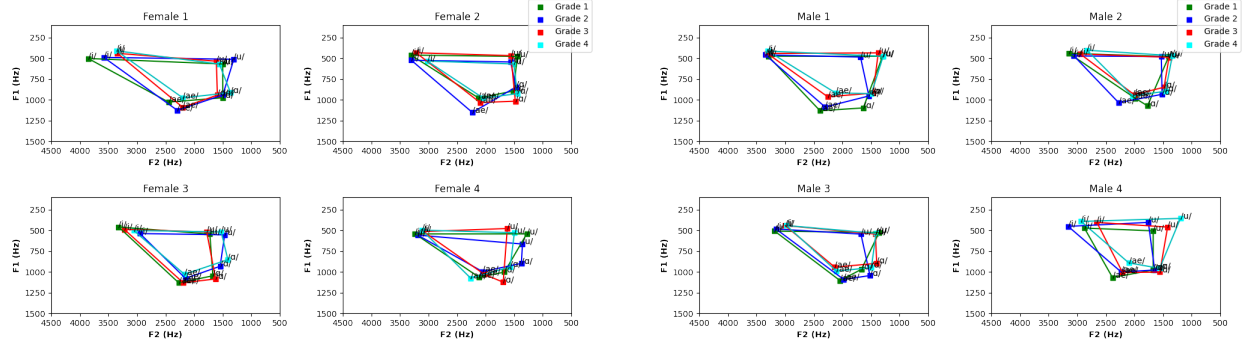


Figure 4.1: Midsagittal tongue profiles during the production of /u/ in “Zoo” and /æ/ in “Vacuum” by a boy in Grade 1. The tongue contour is highlighted using dotted markings.

The articulatory analysis tests two main hypotheses. First, because high vowels require raising the tongue body, TCD is expected to be larger for high vowels than for low vowels. Second, because front vowels require a more anterior tongue configuration, TCP is expected to be larger for front vowels than for back vowels. The analysis therefore compares TCD for high versus low vowels and TCP for front versus back vowels. Additional comparisons examine gender-based differences and grade-based developmental changes.

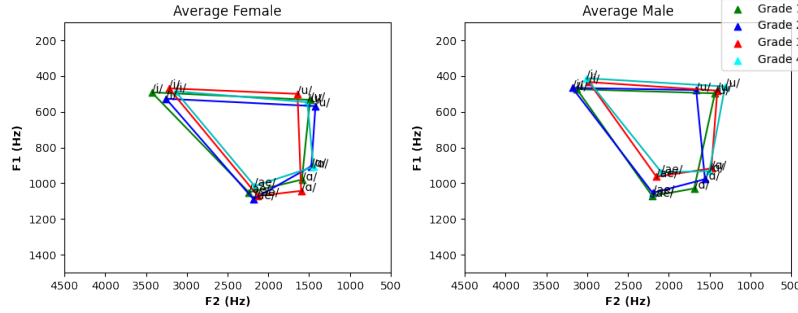
4.2.4 Statistical Comparisons

The results are presented using vowel-space plots, normalized vowel-space plots, articulatory vowel-space plots, and box plots. Statistical comparisons were performed to evaluate differences between vowel classes, gender groups, and grades. Two-sample one-tailed *t*-tests were used for binary comparisons such as high versus low vowels, front versus back vowels, and male versus female children. One-way ANOVA was used to test for grade-based differences in TCD and TCP.



(a) Grade-wise F1–F2 plots for four vowels for four female children.

(b) Grade-wise F1–F2 plots for four vowels for four male children.



(c) Grade-wise averaged F1–F2 plots for four vowels by gender.

Figure 4.2: Longitudinal changes in the first two formant frequencies from Grade 1 to Grade 4. F1 and F2 form the vertical and horizontal axes, respectively.

4.3 Results and Discussion

4.3.1 Longitudinal Changes in the Acoustic Vowel Space

Figures 4.2a and 4.2b show the longitudinal F1–F2 vowel spaces for female and male children, respectively. Each subplot corresponds to one child, and the different colors represent the vowel spaces from different grades. Figure 4.2c shows the corresponding gender-based average vowel spaces across grades.

To quantify the grade-wise changes, the geometric centroid of each quadrilateral in the F1–F2 space was calculated and plotted in Figure 4.3. Since the axes in the F1–F2 space

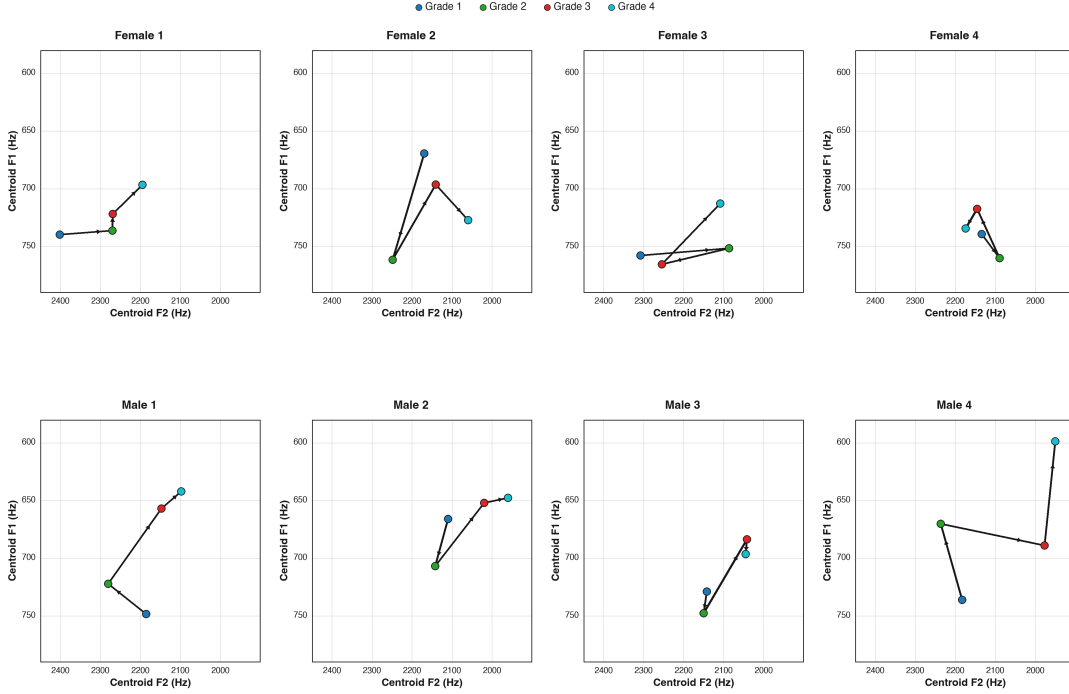


Figure 4.3: Grade-wise changes in the centroids of the vowel spaces for the eight children. The non-monotonic centroid trajectories demonstrate that reductions in formant frequencies with increasing age are not consistent across all children.

are plotted using the conventional vowel-space orientation, a shift up towards the right corresponds to an overall reduction in formant frequency values. Although with grade (i.e., age) vocal-tract growth is generally expected to reduce the formant frequencies values, the centroid trajectories do not decrease consistently (i.e., do not move up towards the right in the F1–F2 space) for all children as seen in Figure 4.3. Several children exhibit increases between adjacent grades, indicating that individual developmental patterns may depart from the expected monotonic trend. These findings demonstrate substantial inter-child and longitudinal variability and suggest that age-related changes in children’s formant frequencies cannot be adequately represented by a single uniformly decreasing trajectory.

This observation is important for the broader goals of this dissertation. In child speaker verification, age-related variation cannot be modeled only as a global age effect. The same grade transition may produce different acoustic changes for different children. Therefore,

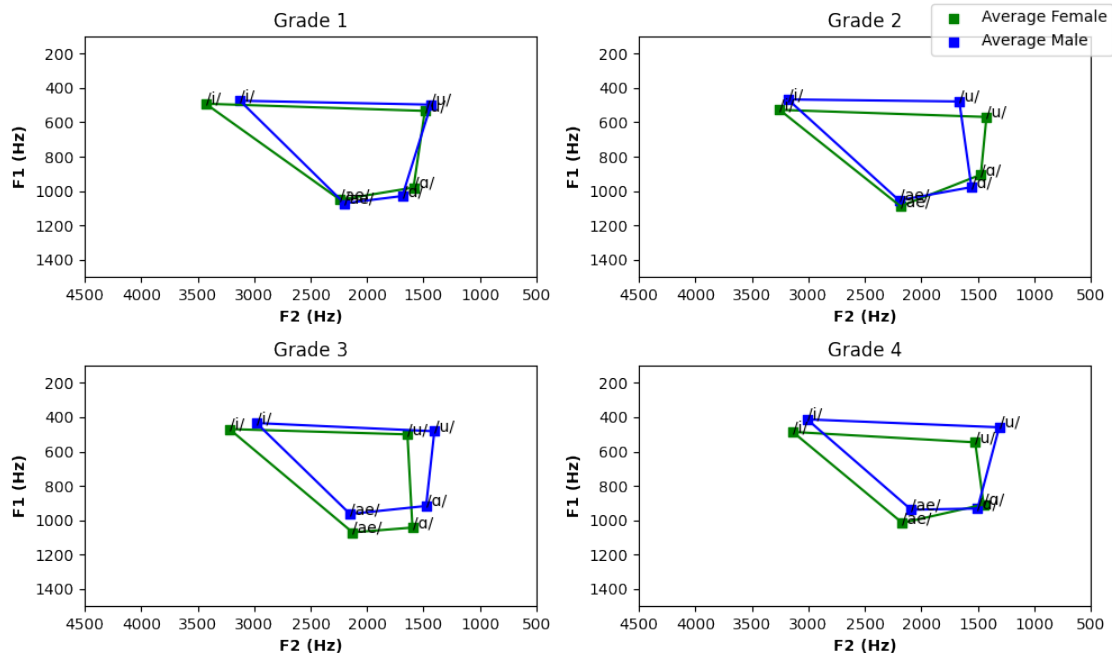


Figure 4.4: Grade-wise average vowel plots highlighting gender-based acoustic differences.

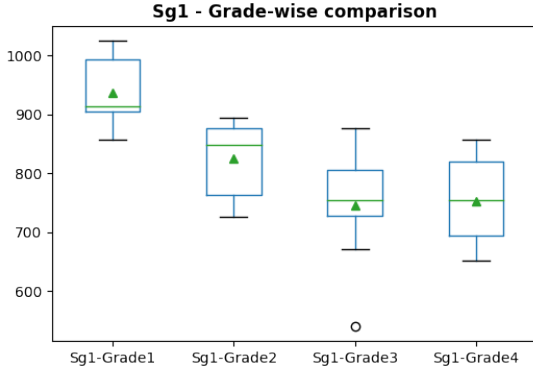
robust child speaker verification systems must handle both systematic developmental trends and child-specific trajectories.

4.3.2 Emergence of Gender-Based Acoustic Differences

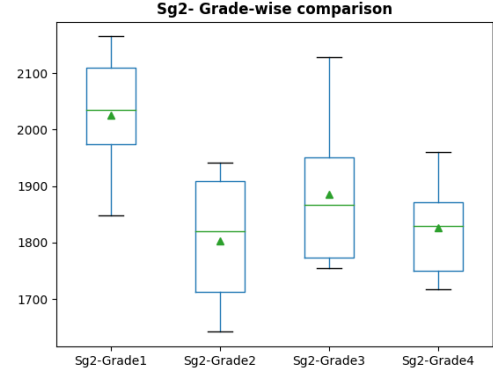
Figure 4.4 shows grade-wise average vowel plots for male and female children. For each grade, formant frequencies were averaged across children of the same gender and plotted as separate male and female vowel spaces.

In Grade 1, the average male and female vowel spaces overlap substantially, suggesting that there are limited gender-based acoustic differences at this age. From Grade 2 onward, the male vowel plots show a shift consistent with lower formant frequency values relative to the female vowel plots. This suggests that gender-based acoustic differences begin to emerge during the elementary school years. This trend is consistent with prior observations of vowel acoustic-space development in children [9].

The emergence of acoustic differences before puberty is relevant because anatomical di-



(a) Grade-wise Sg1 for all children.



(b) Grade-wise Sg2 for all children.

Figure 4.5: Grade-wise distributions of the first two subglottal resonances. The mean value is represented by a green triangle marker within each box plot.

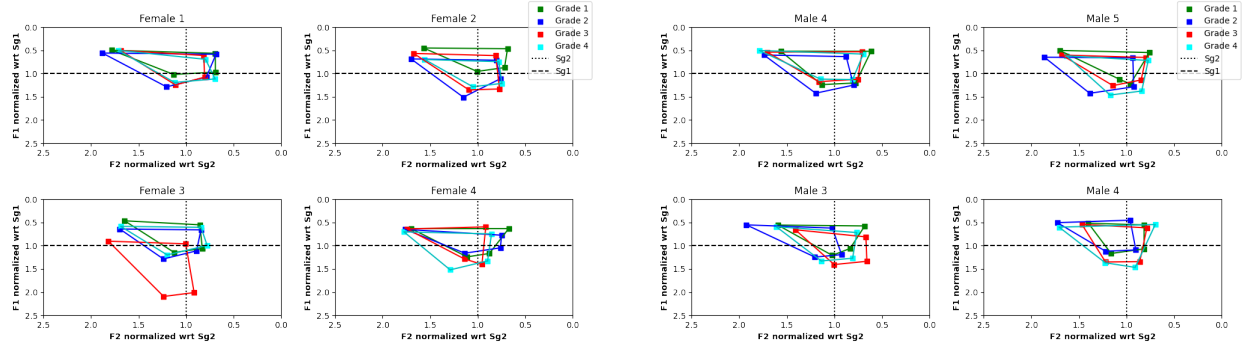
morphism in vocal tract length is generally associated with later development. The results therefore suggest that acoustic differences in children may reflect a combination of anatomical growth, articulatory behavior, and speech motor development. For speaker verification, this further supports the need to understand how age and gender-related variability interact with speaker identity in child speech.

4.3.3 Development of Subglottal Resonances

Figure 4.5 shows the grade-wise distributions of Sg1 and Sg2. Both resonances show an overall reduction from Grade 1 to Grade 4. This decrease is consistent with developmental changes in the subglottal system as children grow.

To examine the role of SGRs in the vowel space, F1 and F2 were normalized using Sg1 and Sg2, respectively. The normalized vowel plots are shown in Figure 4.6. Each subplot corresponds to one child, and the dotted horizontal and vertical lines represent the normalized Sg1 and Sg2 values.

The normalized plots indicate that SGRs provide meaningful reference boundaries for the vowel space. With increasing grade, the SGRs become slightly clearer boundaries separating



(a) Grade-wise SGR-normalized F1–F2 plots
for four female children.

(b) Grade-wise SGR-normalized F1–F2 plots
for four male children.

Figure 4.6: Grade-wise vowel plots with F1 and F2 normalized by the first and second subglottal resonances, respectively. Dotted horizontal and vertical lines indicate normalized subglottal resonance frequencies.

vowel regions, particularly around Grades 2 and 3. This pattern is consistent with prior observations that subglottal resonances structure vowel acoustics in a manner similar to their effect in adult speech [66, 67, 15, 68].

These results show that developmental variability is not limited to supraglottal vowel formants. The subglottal system also changes across grades and interacts with the organization of the vowel space.

4.3.4 Relationship Between Tongue Curvature and Vowel Categories

The articulatory analysis first examines whether TCD and TCP capture expected vowel-category differences. Following the geometric interpretation of these measures, larger TCD values were expected for high vowels, and larger TCP values were expected for front vowels. These hypotheses were tested using the box plots in Figure 4.7.

The results from Figure 4.7a show that TCD is larger for high vowels than for low vowels. This supports the interpretation that TCD is related to vowel height, since high vowels involve raising the tongue body toward the palate. The results in Figure 4.7b also

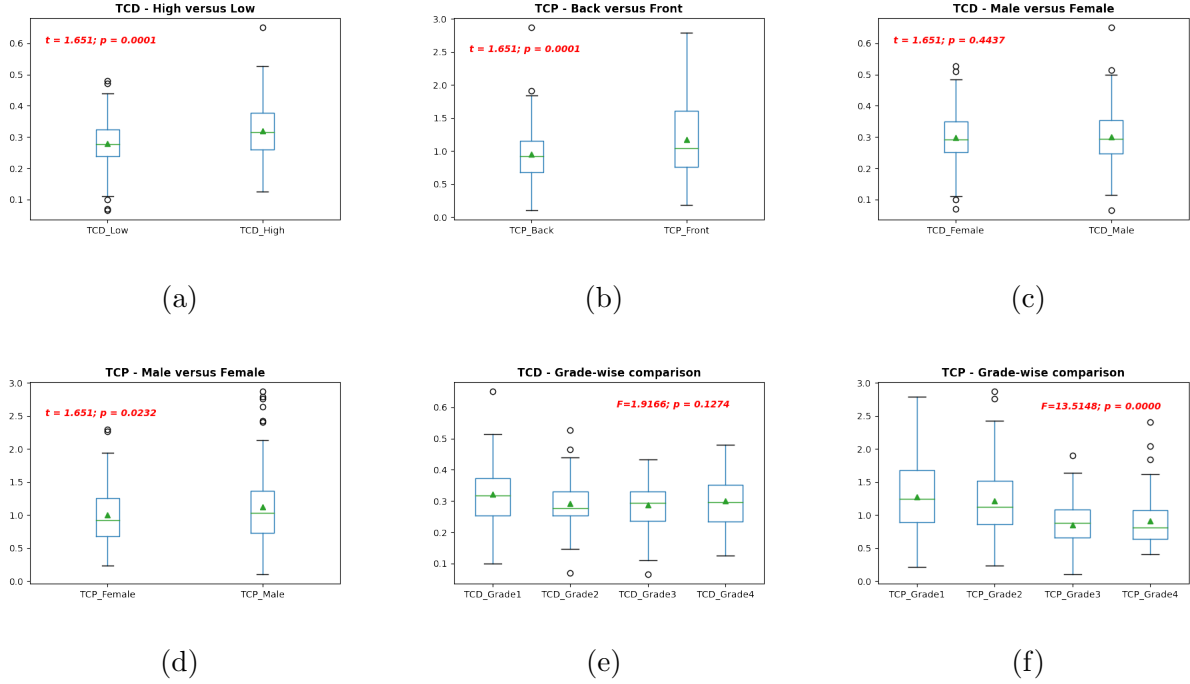


Figure 4.7: Box plots comparing articulatory parameters TCD and TCP. The plots summarize vowel-category, gender-based, and grade-based comparisons. Statistical test results are shown in red in each subplot, and the mean value is represented by a green triangle marker.

show that TCP is larger for front vowels than for back vowels, supporting the interpretation that TCP is related to vowel frontness. These trends indicate that TCD and TCP capture meaningful articulatory information about vowel production.

4.3.5 Gender-Based and Grade-Based Articulatory Patterns

The gender-based TCD and TCP comparisons are shown in Figures 4.7c and 4.7d. In Fig. 4.7c, “TCD_Male” represents TCD values for all vowels /æ/, /i/, /u/, and /ɑ/ over the four years for all male children (128 measurements = 4 vowels $\times$ 2 words per vowel $\times$ 4 male children $\times$ 4 years). Similarly, “TCD_Female” is plotted in Fig. 4.7c and “TCP_Male” and “TCP_Female” in Fig. 4.7d. The mean value is represented by a green triangle marker within each boxplot. The p-value obtained from 2-sample, 1-tailed t-tests comparing TCD and TCP of males and females are given in red font within Figs. 4.7c and 4.7d. Unlike

the acoustic formant results, the articulatory measures do not show significant gender-based differences. This suggests that the observed gender-based acoustic differences are not fully explained by the TCD and TCP measures alone.

The grade-based comparisons are shown in Figures 4.7e and 4.7f. Each boxplot in Fig. 4.7e and 4.7f, was plotted using 64 TCD/TCP measurements (4 vowels $\times$ 2 words per vowel $\times$ 8 children). The grade-wise TCD distributions do not show the same level of systematic change as the grade-wise TCP distributions. In contrast, TCP shows a significant grade effect. The decreasing TCP values across grades may reflect developmental changes in how children coordinate tongue body and tongue blade gestures. One possible interpretation is that, as children grow, they develop more differentiated articulatory control, with tongue blade gestures becoming more reserved for consonantal articulation. This interpretation is consistent with the idea that articulatory strategies adapt during vocal tract growth [69].

The contrast between the acoustic and articulatory results is important. Gender-based differences are apparent in the acoustic vowel space but not in the TCD and TCP measures. Grade-related changes appear in both acoustic and articulatory analyses, but the patterns are not identical. Both must be considered when interpreting developmental variability in child speech.

4.3.6 Articulatory Vowel Spaces Across Grades

Since TCD and TCP are correlated with vowel height and vowel frontness, we hypothesized that the vowel articulatory space using the TCD and TCP features might look similar to the vowel acoustic space using F1 and F2. Figure 4.8 shows the articulatory vowel spaces formed using TCD and TCP. TCD is plotted along the vertical axis, and TCP is plotted along the horizontal axis. These plots allow comparison between articulatory vowel organization and the acoustic vowel spaces defined using F1 and F2 values.

Although some individual articulatory vowel spaces (e.g. “Male 2” grade 2) show broad resemblance to the corresponding acoustic vowel spaces, this relationship is not consistent across children and grades. High vowels generally show larger TCD values than low vowels,

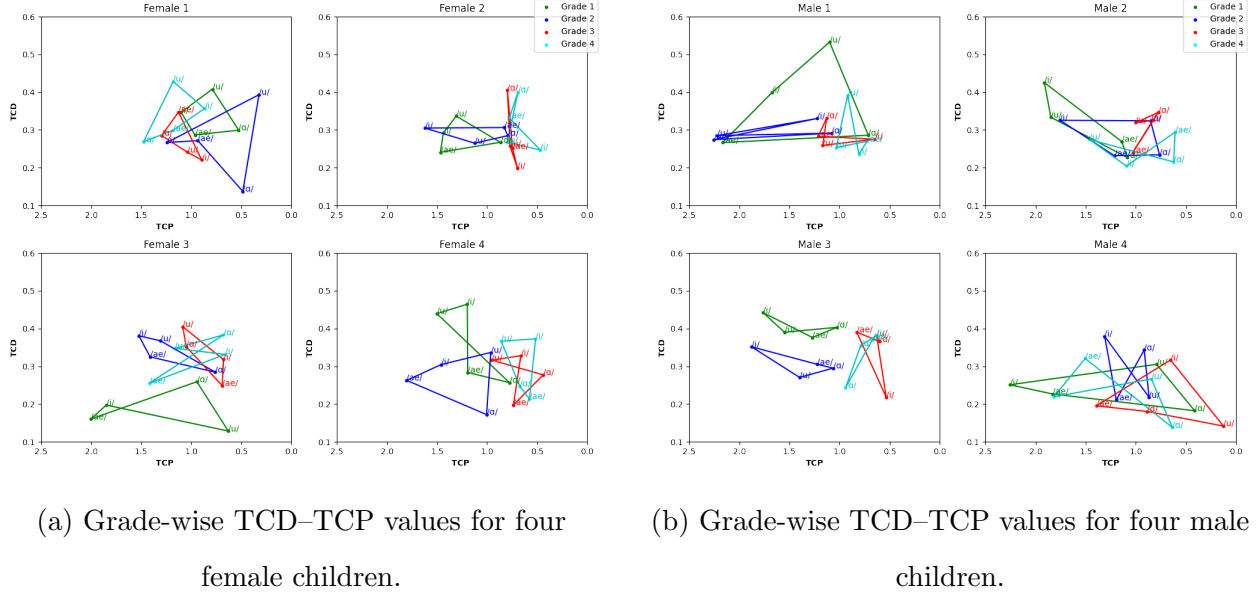


Figure 4.8: Grade-wise TCD versus TCP plots. TCD is plotted along the vertical axis and TCP along the horizontal axis.

but the organization of vowels in the TCD–TCP plane does not always match the organization observed in the F1–F2 plane. This indicates that although TCD and TCP on the whole are correlated with vowel height and frontness, the vowel- and child-specific correlations are not straightforward, especially for TCP. One reason for this could be due to the fact that jaw movement (related to vowel height) and lip rounding (correlated with vowel backness) were not measured.

Another reason could be that the relationship between vowel height and TCD was more straightforward than the relationship between vowel backness and TCP, due to the fact that raising the tongue body (i.e. increasing TCD) creates a vocal tract Helmholtz resonator associated with a low target F1 frequency [70]. In contrast, the target F2 frequency of each vowel is not necessarily uniformly affiliated with either the “front (oral) cavity” or the “back (pharyngeal) cavity” [71]. Specifically, the F2 of /u/ is a second Helmholtz resonance, the F2 of /a/ could be affiliated with either the front or the back cavity depending on their relative lengths, and the F2 of both /i/ and /æ/ is most likely affiliated with the front cavity. This complex relationship between F2 and vocal tract configuration may explain the absence of

a strong resemblance between the vowel acoustic spaces and the corresponding articulatory spaces defined by TCD and TCP.

4.3.7 Implications for Child Speaker Verification

The findings in this chapter directly motivate the speaker verification studies in later chapters. First, the acoustic analysis shows that children’s formant frequencies and subglottal resonances change from Grade 1 to Grade 4. These changes can alter the acoustic characteristics of a child’s speech even when speaker identity remains constant. This creates a challenge for longitudinal child speaker verification, where enrollment and test utterances may be separated by one or more years.

Second, the analysis shows that developmental change varies across children. Some children follow the average trend clearly, while others deviate from it. This means that child speaker verification systems cannot rely only on a simple age-normalization strategy. They must be robust to child-specific developmental trajectories.

As a result, speaker verification systems trained primarily on adult speech may be insensitive to child-specific developmental factors that are not well represented in adult datasets. Overall, this chapter establishes that child speech variability is both developmental and speaker-specific. The following chapters build on this observation by developing age-robust, adult-to-child adapted, and age-agnostic speaker verification methods.

4.4 Summary

This chapter analyzed longitudinal acoustic and articulatory development in children’s speech from Grade 1 to Grade 4. The acoustic analysis showed that F1 and F2 generally decrease across grades, reflecting developmental changes in the vocal tract. Gender-based acoustic differences begin to emerge from Grade 2 onward. The subglottal resonance analysis showed that Sg1 and Sg2 also decrease with age and become useful reference boundaries for organizing the vowel space around Grades 2 and 3.

The articulatory analysis showed that TCD is generally associated with vowel height and TCP with vowel frontness. However, the relationship between acoustic and articulatory vowel spaces is not one-to-one. Children can use different tongue configurations to achieve similar acoustic targets. The analysis did not show significant gender-based differences in TCD or TCP, but TCP showed grade-related developmental change.

The main conclusion is that children’s speech development involves both systematic trends and individual variability. These changes affect the acoustic and articulatory properties of speech and provide an empirical basis for the age-related robustness problem in child speaker verification. The next chapter uses this motivation to study speaker verification systems for intra-year and inter-year child speech evaluation conditions.

CHAPTER 5

Enhancing Age-Related Robustness in Children’s Speaker Verification

5.1 Introduction

The longitudinal analysis in Chapter 4 showed that children’s speech changes across the elementary school years in both acoustic and articulatory dimensions. These developmental changes create a direct challenge for children speaker verification (C-SV). In a practical verification system, a child may enroll at one age and later attempt verification after months or years of growth. In this setting, the system must determine whether the enrollment and test utterances belong to the same child, even though the child’s voice may have changed substantially between the two recordings.

Most speaker verification systems are optimized for speaker-discriminative consistency under relatively stable speaker characteristics [72]. However, in C-SV, age-related variability can alter acoustic cues that are useful for speaker verification. This makes inter-year C-SV more challenging than intra-year C-SV, where enrollment and test utterances are drawn from the same age or grade. While cross-age speaker verification has been studied for adult speech [73], the age-related changes in children’s speech are more pronounced because they are tied to ongoing developmental changes in the vocal tract, articulation, and speech production patterns [6, 74, 75].

In this chapter we will discuss our proposed approaches to improve the robustness of C-SV systems to age-related changes. Specifically, we address two related questions. First, how can C-SV systems be evaluated under longitudinal age-mismatch conditions? Second,

how can the model be trained so that it preserves speaker-discriminative information while reducing sensitivity to local acoustic changes caused by development? To address these questions, we introduce an inter-year C-SV evaluation setup and study two approaches for improving age-related robustness: a Feature Transform Adapter (FTA) and Synthetic Audio Augmentation (SAA).

5.2 Methods

5.2.1 Age-Related Robustness Evaluation

The evaluation of age-related robustness requires verification trials in which enrollment and test utterances may come from different years of the same child’s development. The longitudinal child speech data referred to as IU database used for this purpose were described in Chapter 2. The IU database has child speech data collected longitudinally from eight children from Grade 1 (G1) through Grade 4 (G4).

As discussed in Section 2.3.2, intra-year evaluation trials were constructed by pairing enrollment and test segments from the same grade, producing the G1–G1, G2–G2, G3–G3, and G4–G4 evaluation sets. Inter-year evaluation trials were constructed by pairing enrollment speech from one grade with test speech from a later grade. For example, Grade 1 enrollment utterances were paired with Grade 2, Grade 3, and Grade 4 test utterances, producing G1–G2, G1–G3, and G1–G4 evaluation conditions. Other inter-year sets, such as G2–G3, G2–G4, and G3–G4, were constructed similarly.

For each evaluation set, target trials were formed using utterances from the same child, while non-target trials were formed using utterances from different children. Non-target trials were sampled from the same year to make the task more challenging by reducing the confounding effect of grade mismatch in impostor comparisons. Each evaluation set contains 2,000 verification trials, with 1,000 target and 1,000 non-target trials. This setup allows the evaluation to separately measure intra-year performance and the effect of one-year, two-year, and three-year gaps between enrollment and test speech.

5.2.2 Feature Transform Adapter

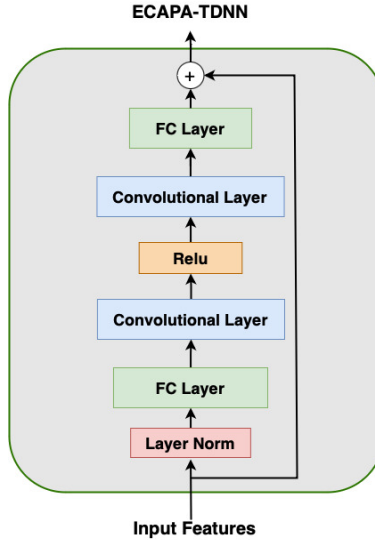


Figure 5.1: Overview of the Feature Transform Adapter used to improve age-related robustness in children speaker verification.

We propose an adapter module called Feature Transform Adapter (FTA), shown in Figure 5.1, to improve robustness to age related variations in SV systems. The motivation for FTA is based on the distinction between local and global speaker information. Local acoustic properties, such as short-term segmental or spectral characteristics, may change as children grow [6]. In contrast, more global speaker characteristics may remain relatively stable and therefore provide more useful cues for speaker verification across years. The goal of FTA is to reduce the model’s dependence on local age-sensitive variation while preserving speaker-discriminative information.

The FTA operates on the 80-dimensional filterbank features used as input to the speaker verification system. The input features are first layer-normalized and passed through a fully connected layer. The resulting features are then processed using two one-dimensional convolutional layers along the frequency dimension, with a ReLU activation between them. These convolutional layers aggregate local frequency-domain information and are intended to smooth or transform local acoustic patterns that may be sensitive to age-related change. The

output of the adapter is then combined with the original input features through a residual connection before being passed to the speaker verification model.

A residual adapter (RA) variant is used as a comparison system. In this variant, the convolutional layers in FTA are replaced by fully connected layers while keeping the number of parameters comparable. This comparison is used to test whether the frequency-domain convolutional operation is important for improving age-related robustness.

5.2.3 Synthetic Audio Augmentation

The second proposed approach is Synthetic Audio Augmentation (SAA). The motivation for SAA is to increase the diversity of the training data while preserving speaker identity. We implement using a pretrained HiFi-GAN generator [76, 77]. For each child speech signal, a spectrogram is provided as input to the generator, which synthesizes a corresponding waveform. The synthetic waveform is then used as additional training data.

The working assumption behind this augmentation is that the synthesis process preserves speaker identity sufficiently for speaker verification training, but may not perfectly preserve fine-grained age-related information. This can be useful for age-robust C-SV because it introduces variability that encourages the model to rely less on age-sensitive local acoustic details. The synthesized samples are therefore used to augment the child speech training data during fine-tuning.

5.2.4 Training and Evaluation Setup

The common speaker verification framework, input features, model architecture, and evaluation metrics were introduced in Chapter 3. Briefly, the systems in this chapter use an ECAPA-TDNN speaker embedding model [2] implemented with the SpeechBrain toolkit [78]. The input consists of 80-dimensional filterbank features extracted using a 25 ms analysis window and a 10 ms frame shift. The model outputs 192-dimensional speaker embeddings, which are compared using similarity scoring for verification.

All systems are evaluated using Equal Error Rate (EER). In addition to the longitudinal IU evaluation sets, the CSLU [49] child speech corpus is used for training and intra-year evaluation. Two train/evaluation splits are used, as described in Chapter 2: CSLU-S, which contains a larger training set, and CSLU-L, which contains fewer training speakers but a larger evaluation set. The CSLU-S split is used for the main age-robustness experiments, while CSLU-L is used to test generalization to a larger intra-year evaluation set and to compare against prior C-SV systems [19].

5.2.5 Models

All models are trained include augmentation approaches discussed in Section 3.5. The models trained only using child speech data (Section 3.6.1) are referred to as *Baseline* systems. Models obtained by finetuning a adult SV system using child speech (Section 3.6.3) are referred to as *Finetune*. In Synthetic Audio Augment experiments we use the pretrained HiFi-GAN vocoder and denoiser [77], with a denoising strength of 0.005, and these models are referred to as *SAA*. The models trained with the Feature Transform Adapter (Section 5.2.2) are referred to as *FTA*. The models trained by replacing convolution layers in the FTA module with fully connected layers are referred to as *RA* (Residual Adapter).

5.3 Results and Discussion

5.3.1 Effect of Feature Transform Adapter

Table 5.1 reports EER results for systems trained using the CSLU-S split and evaluated on CSLU intra-year trials and IU intra-year/inter-year trials. The *Baseline* system performs substantially worse on the IU inter-year conditions than on same-grade, i.e., CSLU Intra-year and IU Intra-year evaluation trials, confirming that age mismatch introduces a significant challenge for C-SV.

Finetune model improves intra-year performance on CSLU and most IU intra-year conditions, but it does not consistently improve inter-year robustness. This suggests that simply

Table 5.1: Equal Error Rate (EER) results for systems trained using the CSLU-S split and evaluated on CSLU intra-year trials (Kindergarten - K00 to Grade 10 - K10) and IU intra-year/inter-year trials. Intra-year refers to enrollment and test segments from the same grade, while inter-year refers to enrollment and test segments from different grades.

<i>CSLU Intra-Year</i>											
	K00	K01	K02	K03	K04	K05	K06	K07	K08	K09	K10
<i>Baseline</i>	13.2	12.85	10.45	9	9.55	7.05	4.7	4.9	4.5	3.45	4.4
<i>Finetune</i>	9.55	10.65	9.05	6.95	7.25	4.1	4.05	2.95	2.75	1.85	3.3
<i>+ RA</i>	9.7	11.5	8.75	7.5	7.1	4.35	3.85	3.2	2.6	1.95	3.45
<i>+ FTA</i>	10.5	10.35	9.05	6.85	5.8	4.2	3.45	2.85	2.55	1.7	3.25
<i>++ SAA</i>	10.6	12.1	9.65	6.8	5.75	4.4	3	2.85	2.6	2.65	3.6

<i>IU Intra-Year</i>					<i>IU Inter-Year</i>					
	G1-G1	G2-G2	G3-G3	G4-G4	G1-G2	G1-G3	G1-G4	G2-G3	G2-G4	G3-G4
<i>Baseline</i>	22.5	15.5	17.5	16.2	29.5	32.8	34.7	39.5	40.3	23.4
<i>Finetune</i>	19	13.9	15.8	11	28.7	36.1	36.7	37.1	39	24.1
<i>+ RA</i>	19.4	16.7	13	13.1	27.9	38	37.1	39.5	38.3	22.3
<i>+ FTA</i>	16.4	12.9	12.2	10.6	26.6	34.2	35.8	34.2	35.2	21
<i>++ SAA</i>	19.9	12.2	8.1	13.1	27.1	31.15	32.6	30.6	32.3	16.8

optimizing the model on child speech is not sufficient to handle longitudinal age mismatch. Because the training data does not explicitly represent all inter-year conditions, the fine-tuned model can still overfit to age-specific or local acoustic patterns.

The *FTA* system improves over conventional fine-tuning on all IU inter-year evaluation sets. This supports the hypothesis that transforming local frequency-domain patterns before speaker embedding extraction can improve robustness to age-related variability. The comparison with the residual adapter (*RA*) further supports this interpretation. Replacing the convolutional layers in *FTA* with fully connected layers does not yield the same level of improvement, indicating that the frequency-domain aggregation performed by the convolutional layers is important for age-related robustness.

5.3.2 Effect of Synthetic Audio Augmentation

Adding *SAA* on top of *FTA* further improves inter-year verification performance. As shown in Table 5.1, the *FTA+SAA* system (i.e., the last row) obtains the lowest EER in five of the six IU inter-year evaluation conditions. Relative to the baseline, the average EER is reduced by 19.4%, 13.0%, and 6.1% for the one-year, two-year, and three-year gap evaluation sets, respectively.

At the same time, *SAA* introduces a trade-off. While it improves inter-year robustness, it does not consistently improve intra-year performance. On the CSLU intra-year evaluation set, *SAA* degrades EER in several age groups compared to *FTA* alone. A similar pattern is observed for some IU intra-year evaluation sets. This suggests that synthetic augmentation encourages the model to become less sensitive to age-specific local acoustic detail, which helps when enrollment and test speech are separated by age, but can reduce performance when those local details are useful for same-year verification. This trade-off is consistent with observations in inter-year adult speaker verification, where improving robustness to age mismatch can come at the cost of same-age verification accuracy [73].

5.3.3 Effect of Training Split and Comparison with Prior C-SV Systems

Table 5.2 summarizes performance across combined CSLU, IU intra-year, and IU inter-year test sets. The first four rows correspond to systems trained using the CSLU-L split, indicated by a [†]. In this setting, the *FTA*[†] only system does not improve over fine-tuning. This is likely because CSLU-L contains substantially fewer training speakers, making the limited size of the training set the dominant constraint. However, adding *SAA*[†] substantially improves performance, with relative EER improvements of 28.3% on the CSLU intra-year set, 29.3% on the IU intra-year set, and 10.8% on the IU inter-year set compared with the corresponding *Baseline*[†].

We also compare the proposed systems with prior C-SV methods from [19]. The comparison highlights two important observations. First, the *Proposed 3/11* model from [19] was

Table 5.2: EER comparisons across combined CSLU, IU intra-year, and IU inter-year evaluation sets. CSLU is obtained by combining K00 to K10 test sets from the CSLU-L evaluation split. IU Intra-Year and IU Inter-Year are obtained by combining the corresponding IU evaluation sets from Table 5.1. [†] indicates training data from CSLU-L, and ^{*} indicates training data from CSLU-S.

	<i>CSLU</i>	<i>IU Intra-Year</i>	<i>IU Inter-Year</i>
<i>Baseline</i> [†]	12.15	27.02	38.7
<i>Finetune</i> [†]	11.1	19.53	35.15
<i>+FTA</i> [†]	11.58	20.30	36.97
<i>++ SAA</i> [†]	8.71	19.1	34.5
<i>+FTA</i> [*]	-	12.97	31.2
<i>++ SAA</i> [*]	-	13.8	28.72
<i>Proposed 3/11</i> [19]	16.77	24.03	41.65
<i>Baseline 3/5</i> [19]	8.15	10.24	30.83

trained using 11 different data augmentation techniques on the out-of-domain dataset Vox-Celeb2 [52]. We observe that incorporating in-domain child speech CSLU data significantly outperforms using only out-of-domain data augmentation for both intra-year and inter-year verification.

Second, conventional training and augmentation methods alone are insufficient for addressing robustness issues in inter-year C-SV, whereas larger datasets of child speech show more pronounced benefits in intra-year verification. Specifically, we observed that the *Baseline 3/5* model [19], trained using substantially more MyST child speech data performs notably better on the IU intra-year evaluation set. This is likely because the MyST data (Grades 3 to 5) used for *Baseline 3/5* aligns better with the IU dataset (Grades 1 to 4). On the other hand, our proposed model trained on 24 hours of CSLU-S data, indicated by ^{*} in Table 5.2, *FTA+SAA*^{*} achieves an IU inter-year EER of 28.72%, outperforming the reported *Baseline 3/5* system that obtained an IU inter-year EER of 30.83%.

5.4 Summary

In this chapter we discussed age-related robustness in children speaker verification. A longitudinal evaluation setup was used to measure both intra-year and inter-year verification performance from Grade 1 to Grade 4. The results showed that inter-year C-SV is substantially more challenging than intra-year C-SV, confirming that developmental changes in children’s voices can degrade verification performance.

To improve robustness, this chapter introduced the Feature Transform Adapter and Synthetic Audio Augmentation. The FTA improved inter-year performance by transforming local frequency-domain features while preserving speaker-discriminative information. SAA further improved inter-year performance by increasing data diversity, although it introduced a trade-off with intra-year accuracy. Experiments with different CSLU train/evaluation splits and comparisons with prior C-SV systems showed that explicitly addressing age-related variability is important for robust longitudinal C-SV.

The main conclusion is that age-related robustness should be treated as a distinct objective in child speaker verification. Improving performance on child speech in general is not sufficient; systems must also be designed to handle the developmental changes that occur within the same child over time.

CHAPTER 6

Gated Linear Unit Adapter with Iterative Fine-Tuning

6.1 Introduction

The previous chapter focused on improving age-related robustness within children speaker verification (C-SV), especially when enrollment and test utterances are separated by one or more years. This chapter addresses a related but broader problem: how to efficiently adapt speaker verification systems trained on high-resource adult speech to the low-resource child speech domain.

Modern speaker verification systems trained on adult speech achieve strong performance on large-scale adult datasets, but their performance degrades when directly applied to children’s speech. This degradation is caused by the acoustic mismatch between adult and child speech, including differences in vocal tract size, pitch, formant structure, speaking style, and developmental variability. At the same time, training a high-capacity neural speaker verification system directly on child speech is difficult because child speech datasets are much smaller than adult speech datasets. Therefore, adult-trained models provide a useful starting point, but they must be adapted carefully to the child domain.

A common approach is to fine-tune an adult-trained speaker embedding model using child speech. However, in low-resource C-SV, conventional fine-tuning may not use the limited target-domain data efficiently. In speaker verification, fine-tuning is also different from tasks such as automatic speech recognition (ASR) because the classifier must be changed to match the speakers in the target-domain training set. This means that the classifier is typically trained from scratch, while the pretrained embedding model is adapted from the source domain. Updating all components simultaneously can be inefficient, particularly when the

target-domain dataset is small.

This chapter proposes a Gated Linear Unit adapter with Iterative Fine-Tuning, referred to as G-IFT, for efficient adult-to-child adaptation. The goal is to improve knowledge transfer from adult-trained speaker verification models to child speech while limiting overfitting in low-resource settings. As discussed in section 3.4, the proposed framework is evaluated across multiple speaker embedding architectures, including ECAPA-TDNN [2], ResNet [3], and X-vector [1] systems, to test whether the approach is architecture-agnostic.

6.2 Methods

6.2.1 Adult-to-Child Adaptation Framework

The adaptation setting follows the general framework introduced in Section 3.6. A speaker embedding model trained on adult speech is adapted to perform C-SV using child speech data. The adapted model is evaluated on child speaker verification trials. The child speech datasets, feature extraction settings, and evaluation metrics used in this chapter were described in Chapters 2 and 3.

The experimental comparison uses several systems. The *Baseline* system refers to a model trained from scratch using child speech (Section 3.6.1). The *Pretrained* system refers to an adult-trained speaker verification model evaluated directly on child speech without adaptation (Section 3.6.2). The *Finetune* system refers to conventional fine-tuning of the adult-trained model using child speech (Section 3.6.3). Residual Adapter (RA) fine-tuning is included as an adapter-based comparison system [79]. The proposed systems include a GLU adapter without iterative fine-tuning and two variants of the full G-IFT framework.

6.2.2 Gated Linear Unit Adapter

The first component of the proposed framework is a Gated Linear Unit (GLU) adapter. The GLU mechanism was originally introduced to allow networks to control information flow

through learned gates [80]. In this chapter, the motivation for using a GLU adapter is to allow the adapted system to learn how much information from the pretrained adult speaker embedding model should be passed forward for child speaker verification.

As shown in Figure 6.1(a), the GLU adapter is inserted between the output of the speaker embedding model and the classifier. Let E_{out} denote the output of the pretrained speaker embedding model. This representation is first passed through a linear layer, followed by a ReLU activation and layer normalization. The resulting representation, denoted G_{in} , is passed through the GLU layer. The operation of the GLU layer is given by

$$G_{\text{out}} = (G_{\text{in}} * W + b) \otimes \sigma(G_{\text{in}} * V + c), \quad (6.1)$$

where W , V , b , and c are learnable parameters, $\sigma(\cdot)$ is the sigmoid activation function, and $\otimes$ denotes element-wise multiplication. The sigmoid branch acts as a learned gate that controls how much information is passed through the adapter. The output of the GLU layer is then passed through another linear layer before being used by the classifier. During evaluation, the output of the adapter is used as the speaker embedding.

6.2.3 Iterative Fine-Tuning - IFT

The second component of the proposed framework is iterative fine-tuning (IFT). Conventional fine-tuning updates the embedding model and classifier simultaneously. In adapter-based fine-tuning, the adapter and classifier are often updated together, and in some cases the embedding model is also updated at the same time. However, as we briefly discussed in Section 6.1, in the case of adaptation unlike ASR, in SV, the classifier has to be initialized from scratch to match the target domain speaker set. Hence classifier and adapter (also initialized from scratch) may require larger adjustments than the pretrained embedding model. Updating all components simultaneously can therefore lead to inefficient optimization, especially when child speech data is limited.

G-IFT addresses this issue by separating the adaptation process into focused update

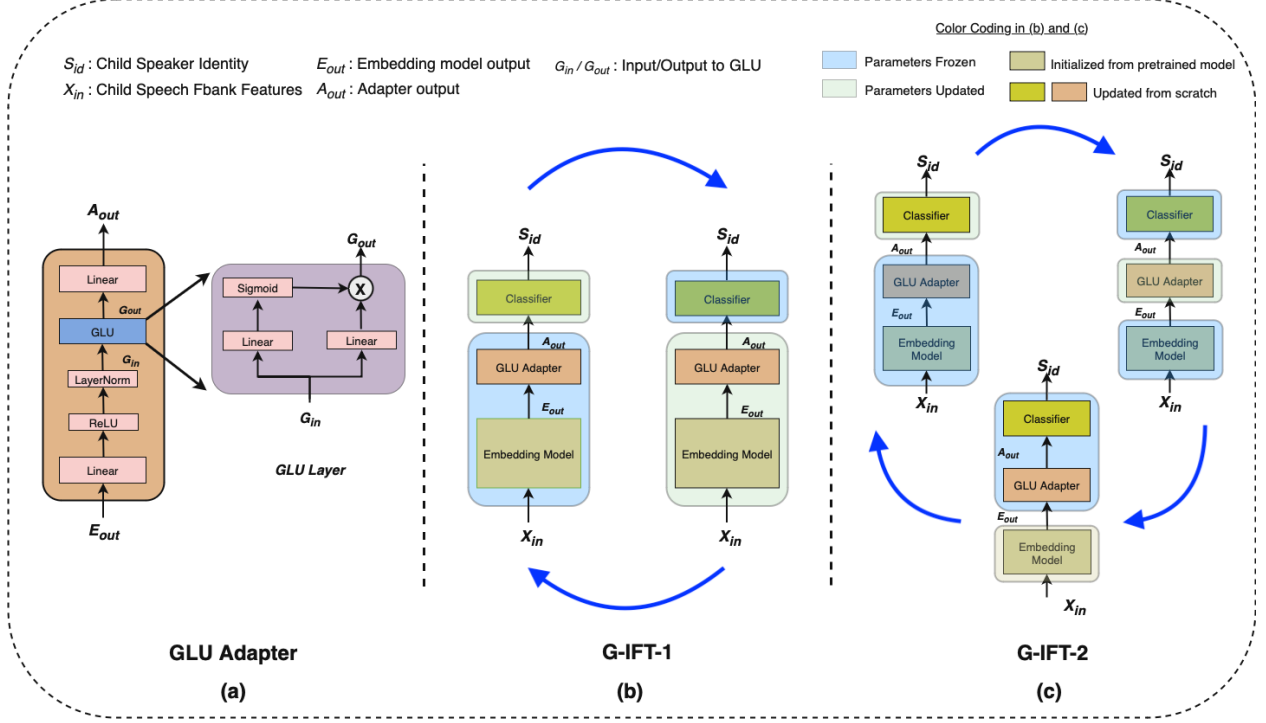


Figure 6.1: Overview of the G-IFT framework. Panel (a) shows the GLU adapter. Panels (b) and (c) show the iterative fine-tuning procedures for G-IFT-1 and G-IFT-2, respectively.

steps. The framework contains two variants: G-IFT-1 and G-IFT-2. Both variants begin by initializing the speaker embedding model with adult-trained parameters and inserting the GLU adapter between the embedding model and the target-domain classifier.

In G-IFT-1 (Figure 6.1(b)), training alternates between two update stages. First, the GLU adapter and classifier are fine-tuned jointly using the child speech data, while the pre-trained speaker embedding model is kept frozen. In the second stage, the pretrained speaker embedding model is finetuned, while the adapter and classifier are kept frozen. This process allowing the adapter and classifier to first adjust to the target-domain speakers before making any updates to the larger embedding model.

In G-IFT-2 (Figure 6.1(c)), the update process is further separated into three stages. The classifier is fine-tuned first, followed by the GLU adapter, and then the pretrained speaker embedding model. This variant allows the classifier to adapt in isolation before adapter and embedding-model updates are introduced. The motivation is that, in extremely low-resource

conditions, separately updating the classifier, adapter, and embedding model may improve the use of limited target-domain data.

To make the comparison fair, the number of training epochs is adjusted so that the number of parameter updates is comparable across fine-tuning methods. The total number of epochs for G-IFT-1 is twice that of conventional fine-tuning, while the total number of epochs for G-IFT-2 is three times that of conventional fine-tuning. This accounts for the fact that G-IFT-1 and G-IFT-2 split updates across multiple stages rather than updating all components together in each epoch.

6.2.4 Experimental Comparison

The experiments evaluate whether the proposed framework improves adult-to-child adaptation under different target-domain data conditions. The framework is tested using ECAPA-TDNN [2], ResNet [3], and X-vector [1] speaker verification architectures. All systems are implemented using the SpeechBrain toolkit [78]. The adult pretrained models are trained using VoxCeleb [51], while the child-domain adaptation and evaluation are conducted using CSLU [49] and MyST [50].

The MyST experiments use multiple target-domain training subsets with increasing amounts of train data as discussed in Section 2.3.3. These subsets allow the analysis to test whether the advantage of G-IFT depends on the amount of available child speech data. This is important because the main motivation for the framework is low-resource adaptation. If the method is effective, the largest improvements should appear when the amount of child speech adaptation data is small. All systems are evaluated using Equal Error Rate (EER).

6.3 Results and Discussion

6.3.1 Performance Across Speaker Verification Architectures

Table 6.1 compares the proposed methods with pretrained, baseline, conventional fine-tuning, residual adapter, and GLU adapter systems on CSLU and MyST-1. The results show that directly applying adult pretrained models to child speech, i.e., *Pretrained* is not sufficient, confirming the adult-to-child domain mismatch. Training from scratch on child speech, i.e., *Baseline* is competitive in some cases, but it is not consistently better than conventional fine-tuning, i.e., *Finetune*, especially when the child training data is limited.

Finetune improves performance over the pretrained system in most cases, showing that adult-trained speaker embeddings provide useful initialization for C-SV. However, the proposed adapter-based adaptation, i.e., *GLU* provides additional gains over conventional fine-tuning in four of the six architecture/dataset combinations, indicating that gated adaptation can help the model selectively transfer information from the adult-trained embedding model to the child-domain task.

The full G-IFT framework provides the most consistent improvement. Across ECAPA-TDNN, ResNet, and X-vector systems, either *G-IFT-1* or *G-IFT-2* achieves the best performance for every architecture/dataset pair. This supports the central hypothesis of the chapter: the benefit of adapter-based adaptation is strengthened when the classifier, adapter, and pretrained embedding model are optimized through an iterative fine-tuning strategy rather than updated simultaneously.

The results also show that the best G-IFT variant depends on the architecture and dataset. For ECAPA-TDNN and ResNet on CSLU, G-IFT-1 performs best. For ECAPA-TDNN and ResNet on MyST-1, and for X-vector on both CSLU and MyST-1, G-IFT-2 performs best. This suggests that separating the classifier, adapter, and embedding model updates can be particularly useful in some low-resource adaptation settings, but the best update schedule may depend on model architecture and target-domain data characteristics.

In order to verify the improvement brought out by the proposed GLU adapter, an addi-

Table 6.1: Performance in terms of Equal Error Rate (EER %) across CSLU and MyST-1. *Pretrained* refers to adult SV systems tested directly on child speech. *Baseline* refers to systems trained from scratch on child speech. *Finetune* refers to conventional fine-tuning of adult SV systems using child speech. *GLU* refers to adapter fine-tuning, and G-IFT-1/G-IFT-2 are the proposed iterative fine-tuning frameworks. Statistical significance is assessed using paired t -tests at $p = 0.05$ and significant improvements over Finetune are marked with “*”.

	ECAPA-TDNN		ResNet		X-vector	
	CSLU	MyST-1	CSLU	MyST-1	CSLU	MyST-1
<i>Pretrained</i>	17.39	17.48	12.62	14.47	23.59	24.95
<i>Baseline</i>	12.15	24.80	14.23	36.23	14.71	25.70
<i>Finetune</i>	11.10	20.04	7.45	10.64	15.89	21.26
<i>RA</i> [79]	11.34	22.89	7.47	12.76	15.61	21.54
<i>GLU</i>	9.70	22.30	6.92	12.48	12.31	17.96
<i>G-IFT-1</i>	8.88*	16.38	6.77*	9.29	12.71	18.20
<i>G-IFT-2</i>	9.06	14.22*	6.88	8.37*	12.09*	16.79*

tional comparison is obtained by applying IFT without the GLU adapter. On MyST-1, IFT alone gives EERs of 22.5%, 10.2%, and 20.9% for ECAPA-TDNN, ResNet, and X-vector, respectively. In contrast, as seen in Table 6.1, applying IFT with the GLU adapter, i.e., *G-IFT-1*, *G-IFT-2*, reduces the corresponding EERs to 14.2%, 8.4%, and 16.8%. This indicates that the improvement does not come from iteration alone. Rather, the combination of the adapter-classifier structure and the iterative update strategy provides the benefit. And as far as comparison with other adapter approaches is concerned, although RA performed well in [79], we can see from the results in Table 6.1 that it is less effective here.

Table 6.2: Equal Error Rate (EER %) across ECAPA-TDNN, ResNet, and x-vector systems on different MyST training splits. The duration of the training splits increases progressively from MyST-1 to MyST-4. The evaluation split is fixed, with no overlap in speakers between training and evaluation sets. Statistical significance is assessed using paired t -tests at $p = 0.05$ and significant improvements over Finetune are marked with “*”.

	ECAPA-TDNN				ResNet				X-vector			
	MyST-1	MyST-2	MyST-3	MyST-4	MyST-1	MyST-2	MyST-3	MyST-4	MyST-1	MyST-2	MyST-3	MyST-4
Baseline	24.80	14.42	10.76	10.11	36.23	19.26	18.05	11.48	25.70	19.76	17.08	16.24
Finetune	20.04	7.97	5.81	6.05	10.64	9.33	6.60	4.87	21.26	19.16	16.70	16.11
RA [79]	22.89	8.19	6.13	6.07	12.76	8.62	5.70	4.87	21.54	19.50	16.36	17.83
GLU	22.30	7.51	5.92	5.97	12.48	7.48	5.50	5.07	17.96	15.05	11.49*	11.75*
G-IFT-1	16.38	7.03	5.49*	5.42*	9.29	6.42	4.89*	4.82	18.20	15.37	12.29	12.58
G-IFT-2	14.22*	7.81	5.93	5.99	8.37*	5.87*	5.52	5.27	16.79*	14.58*	14.70	15.30

6.3.2 Effect of Target-Domain Training Data Size

Table 6.2 reports EER results across the MyST-1 to MyST-4 training subsets. The amount of child speech data increases from MyST-1 to MyST-4 while the evaluation set remains fixed. This setup allows the analysis to separate low-resource and higher-resource adaptation behavior.

Across all methods, EER generally decreases as the amount of MyST training data increases. This is expected: more in-domain child speech gives the model more opportunity to learn child-domain speaker-discriminative patterns. However, the effect of the proposed G-IFT framework is strongest in the low-resource conditions. On MyST-1 and MyST-2, G-IFT-1 and G-IFT-2 consistently improve over conventional fine-tuning across all three architectures. This confirms that the proposed method is especially useful when child speech data is scarce.

Figure 6.2 summarizes the absolute EER reduction of the proposed G-IFT framework over conventional fine-tuning. The largest and most consistent reductions occur in MyST-1 and MyST-2. These are the low-resource conditions where the model is most vulnerable to overfitting and inefficient adaptation. By updating the classifier, adapter, and embed-

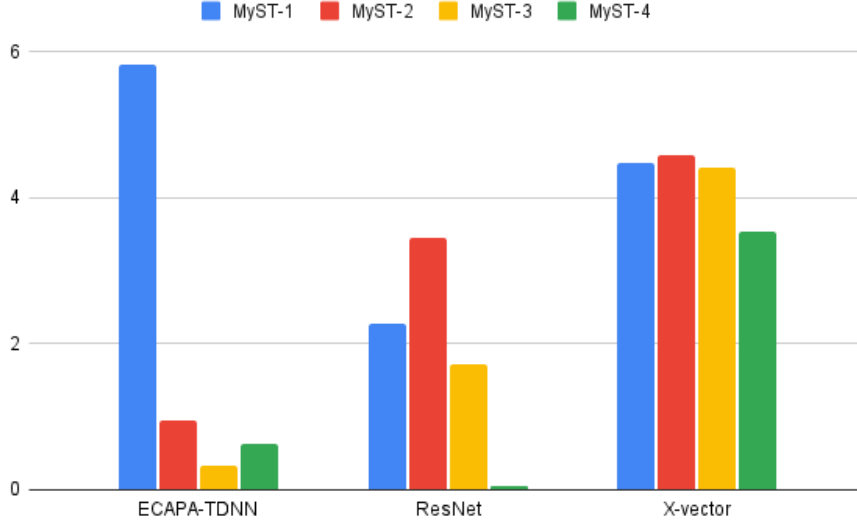


Figure 6.2: Absolute reduction in EER (%) by the proposed G-IFT framework (we consider the best result out of G-IFT-1 and G-IFT-2) over conventional fine-tuning across different MyST training splits. Larger reductions are observed in low-resource settings, especially MyST-1 and MyST-2.

ding model in a staged manner, G-IFT appears to use the limited child speech data more effectively.

The advantage of G-IFT becomes smaller in higher-resource conditions. On MyST-3 and MyST-4, conventional fine-tuning becomes more competitive, especially for ECAPA-TDNN and ResNet. This suggests that when sufficient child speech data is available, the model can adapt effectively even without the staged update strategy. In contrast, when only a small amount of child speech is available, the structure imposed by G-IFT provides a more efficient adaptation path.

The relative behavior of G-IFT-1 and G-IFT-2 also changes with training data size. In the lower-resource MyST-1 and MyST-2 conditions, G-IFT-2 often performs better, suggesting that separately updating the classifier, adapter, and embedding model can be useful when target data are scarce. In the larger MyST-3 and MyST-4 conditions, G-IFT-1 is often more competitive, suggesting that jointly updating the adapter and classifier may be sufficient

when more target-domain data are available.

6.3.3 Discussion

The results in this chapter show that adult-to-child adaptation is not simply a matter of fine-tuning all parameters on child speech. Because child speech data is limited, the adaptation procedure itself affects how efficiently adult-domain knowledge is transferred. The proposed GLU adapter introduces a gated mechanism that learns how much information should pass from the pretrained embedding model to the child-domain classifier. The iterative fine-tuning strategy then structures the optimization process so that the classifier and adapter can adapt before larger changes are made to the pretrained embedding model.

The architecture-agnostic behavior of the proposed framework is also important. G-IFT improves performance across ECAPA-TDNN, ResNet, and X-vector systems, which suggests that the method is not tied to a single speaker embedding architecture. This is useful for the broader dissertation because the goal is not only to improve one particular model, but to study general adaptation strategies for C-SV.

The low-resource results are especially relevant to practical C-SV. Child speech data is often expensive to collect, difficult to annotate, and limited by privacy and consent constraints. A method that improves adaptation using small amounts of child speech is therefore valuable. The MyST subset experiments show that G-IFT provides its largest advantage precisely in these low-resource conditions.

This chapter also sets up the next stage of the dissertation. The methods here improve performance on child speech by adapting adult-trained models to the child domain. However, adapting a model to child speech can degrade its performance on the original adult domain. This trade-off leads to the catastrophic forgetting problem studied in the following chapters, where the goal is to build systems that improve C-SV while retaining A-SV performance.

6.4 Summary

Experiments on CSLU and MyST showed that the proposed framework consistently improves over conventional fine-tuning across ECAPA-TDNN, ResNet, and X-vector architectures. The largest improvements were observed in low-resource MyST conditions, confirming that G-IFT is particularly effective when only limited child speech data is available. As the amount of target-domain data increases, conventional fine-tuning becomes more competitive, but G-IFT remains a useful adaptation strategy.

The main conclusion of this chapter is that efficient adaptation is critical for C-SV. A-SV models contain useful speaker-discriminative knowledge, but this knowledge must be transferred carefully to the child domain. The next chapter builds on this observation by addressing the source-domain retention problem: how to improve C-SV performance without sacrificing A-SV performance.

CHAPTER 7

Age-Agnostic Speaker Verification

7.1 Introduction

The previous chapter addressed efficient adult-to-child adaptation for low-resource children speaker verification (C-SV). The central goal there was to improve performance on child speech by adapting speaker verification systems trained on adult speech. However, improving target-domain performance alone is not sufficient for building inclusive speaker verification systems. A model adapted to child speech can substantially degrade on the original adult speech domain, even when child-domain performance improves. This behavior is an instance of catastrophic forgetting: the model learns the target domain, but loses part of the knowledge that made it effective on the source domain.

In this chapter we address the next step in the dissertation: building a speaker verification system that is robust across both adults’ and children’s speech. This goal is referred to here as *age-agnostic speaker verification* (AASV). In this setting, the system should perform well on C-SV tasks without sacrificing performance on A-SV tasks. This is important because real-world voice-based systems are not always deployed for a single age group. A practical speaker verification system may need to handle both adults and children, and therefore should not require a strict choice between adult-domain and child-domain performance.

The challenge arises because adult and child speech differ acoustically due to developmental differences in vocal tract structure, pitch, formant patterns, and speaking behavior [81, 82, 83]. Adult-trained speaker verification systems perform strongly on adult benchmark datasets, but often perform poorly on children speech. Conversely, child-adapted systems improve C-SV performance, but may degrade A-SV performance. Therefore, the central

problem addressed in this chapter is: *how can adult- and child-domain speaker information be combined to form a unified speaker representation that performs well across age groups?*

Unlike the source-free setting studied in the next chapter, this chapter assumes that source-domain information is available when constructing the age-agnostic system. This allows the system to use both adult-domain and child-domain speaker embedding models, and to learn age-domain information that guides the integration of their embeddings. The proposed solution is an Age-Agnostic Speaker Verification (AASV) system that combines adult- and child-specific speaker representations using a domain classifier. The goal is to reduce the adult-child performance trade-off and build a more inclusive speaker verification system.

7.2 Methods

7.2.1 Age-Agnostic Speaker Verification (AASV) Framework

The proposed AASV system combines information from two domain-specific speaker verification systems: an adult speaker verification system and a child speaker verification system. The adult system is trained on adult speech and is effective for A-SV, while the child system is obtained by adapting the adult system to child speech and is effective for C-SV. Rather than choosing one system for all inputs, the proposed framework uses age-domain information to determine how adult- and child-domain embeddings should be combined.

Figure 7.1 shows the workflow of the proposed AASV system. Given an input utterance, embeddings are extracted from both the adult-domain and child-domain speaker embedding networks. A domain classifier estimates whether the input speech is more likely to belong to the adult or child domain. The domain classifier output is then used to weight and concatenate the two embeddings, producing an expanded age-agnostic representation. This representation is used for speaker verification scoring.

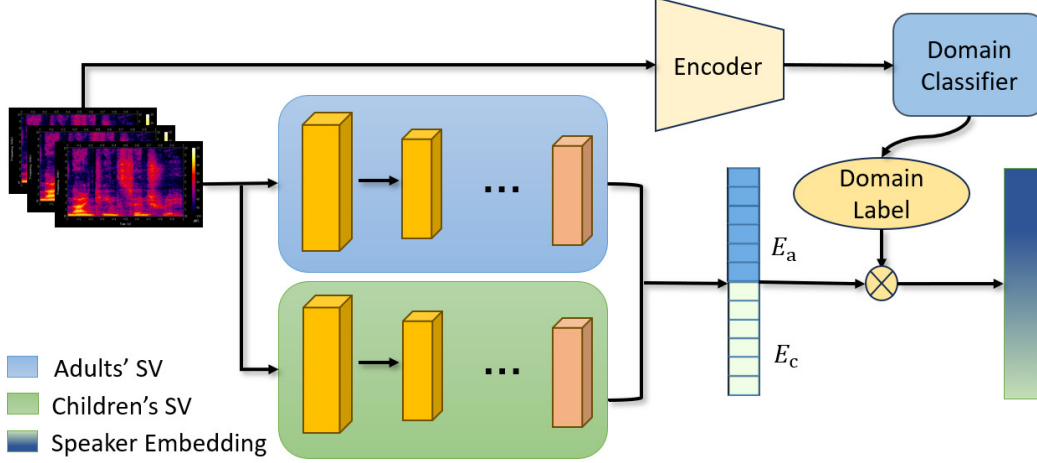


Figure 7.1: Workflow of the Age-Agnostic Speaker Verification (AASV) system. Speaker embeddings E_a and E_c from the A-SV and C-SV subsystems are fused with age-domain information from the domain classifier to generate robust speaker embeddings.

7.2.2 Domain-Specific Speaker Embedding Networks

The first component of the AASV framework is a pair of domain-specific speaker embedding networks. The adult-domain network is trained on adult speech and maps speech inputs to adult-domain speaker embeddings. Let this network be parameterized by θ_a , with classification head parameters defined by adult speaker set. The resulting adult-domain embedding network extracts speaker embeddings E_a that are well suited to adult speech. It is to be noted in all our experiments, we do not train A-SV models, instead use existing pretrained models.

To obtain the child-domain network, the adult-domain network is fine-tuned using child speech. The speaker embedding network is initialized with the adult-domain parameters and adapted to the child domain, yielding parameters θ_c . A new classification head is trained with parameters defined by the child-domain speakers. The resulting child-domain network extracts speaker embeddings E_c that are better suited to C-SV.

Maintaining separate adult- and child-domain embedding networks has two advantages. First, it avoids forcing a single embedding network to represent both domains under strong

data imbalance. Adult datasets are typically larger and more diverse than child datasets, so training a single mixed-domain model may bias the representation toward adult speech. Second, separate networks make it possible to preserve adult-domain speaker information while adapting a child-domain model. The age-agnostic system then combines the two representations instead of replacing one with the other.

7.2.3 Age-Domain Representation and Domain Classifier

The second component of the AASV framework is a domain classifier that estimates age-domain information from speech. The motivation for this component is that speaker embeddings contain information related not only to speaker identity but also to age-domain characteristics. Prior work has shown that speaker embeddings can carry age-related information [31]. In this chapter, this property is used to guide the integration of adult- and child-domain embeddings.

A pretrained ECAPA-TDNN model [2] is used as an encoder to convert speech samples of varying duration into fixed-dimensional embeddings. Figure 7.2 shows t-SNE visualizations of such fixed-dimensional embeddings from CSLU child speech and VoxCeleb adult speech. A domain classifier is then trained to distinguish between adult and child speech using these embeddings.

7.2.4 Multi-Space Embedding Integration

The third component of the AASV framework is multi-space embedding integration. For each input utterance, the adult-domain network produces an embedding $\mathbf{E}_a \in \mathbb{R}^d$, and the child-domain network produces an embedding $\mathbf{E}_c \in \mathbb{R}^d$. The domain classifier produces a softmax output $\mathbf{p} = [p_c, p_a]$, where p_c is the estimated probability that the input belongs to the child domain and p_a is the estimated probability that it belongs to the adult domain.

The domain probabilities are used as dynamic fusion coefficients. The child-domain embedding is weighted by p_c , and the adult-domain embedding is weighted by p_a . The two

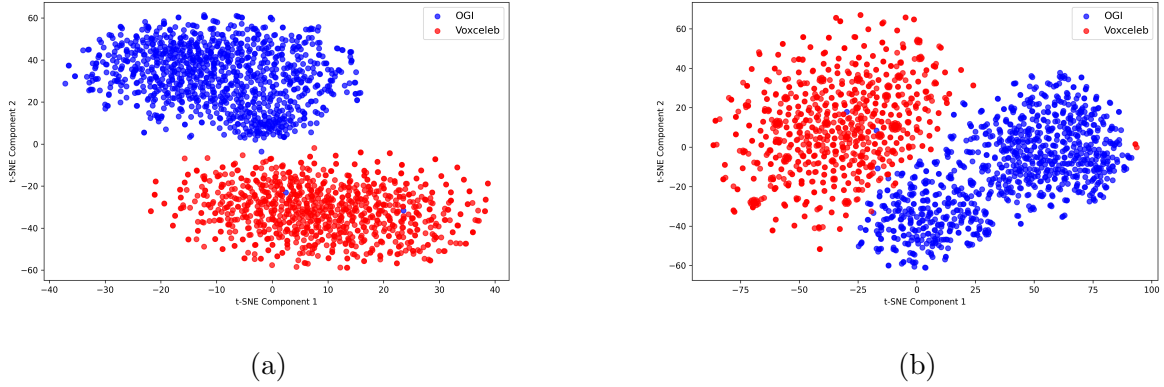


Figure 7.2: t-SNE visualization of speaker embeddings. Blue and red points denote children’s speech from CSLU and adult speech from VoxCeleb, respectively. (a) Younger children’s group (K00–K06) compared with adult speech. (b) Older children’s group (K07–K10) compared with adult speech.

weighted embeddings are concatenated to form the age-agnostic representation:

$$\mathbf{E}_R = [p_c \mathbf{E}_c; p_a \mathbf{E}_a] \in \mathbb{R}^{2d}. \quad (7.1)$$

This expanded representation preserves information from both embedding spaces while allowing the domain classifier to control their relative contribution. If an input is clearly child-like, the representation emphasizes the child-domain embedding. If it is clearly adult-like, it emphasizes the adult-domain embedding. If the age domain is uncertain, both embeddings contribute, which can reduce the impact of incorrect hard domain assignment.

7.2.5 Experimental Comparison

The experiments compare adult-domain, child-domain, and age-agnostic systems. The adult-domain systems are referred to as A-SV systems, and the child-domain systems are referred to as C-SV systems. Two ECAPA-TDNN configurations are used: a smaller model with 512 channels per convolutional frame layer and a larger model with 1024 channels per convolutional frame layer. These are referred to as the Small and Large systems, respectively.

Table 7.1: Accuracy of domain classifiers trained on mixed datasets with varying proportions of adult and child speech. The children test set is a mixture of K00–K10 CSLU test sets, and the adult test set is VoxCeleb1-O.

Adult:Child Ratio	# Utterances		Accuracy	
	Adults	Children	Adults	Children
<i>1:1</i>	20,000	20,000	42.1%	100.0%
<i>2:1</i>	40,000	20,000	70.6%	99.9%
<i>3:1</i>	60,000	20,000	87.5%	99.9%
<i>4:1</i>	80,000	20,000	89.7%	99.7%
<i>5:1</i>	100,000	20,000	95.0%	99.6%

The adult-domain systems are evaluated on VoxCeleb test sets, while the child-domain systems are evaluated on CSLU year-wise test sets. The AASV systems are evaluated on both adults’ and children’s speech to determine whether they can maintain balanced performance across domains. The comparison also includes existing approaches such as ChildAugment [19], Stable Learning [84], and Weight Space Ensemble (WSE) [29]. Equal Error Rate (EER) is used as the evaluation metric.

7.3 Results and Discussion

7.3.1 Domain Classifier Behavior

The first set of experiments examines the domain classifier. Table 7.1 reports the accuracy of classifiers trained with different adult-to-child data ratios. The amount of child speech is fixed at 20,000 utterances, while the amount of adult speech is increased from 20,000 to 100,000 utterances.

The classifier achieves nearly perfect accuracy on the child test set across all data ratios.

In contrast, adult-domain accuracy is poor when equal numbers of adult and child utterances are used, but improves substantially as the amount of adult training data increases. With a 5:1 adult-to-child ratio, adult-domain accuracy reaches 95.0%, while child-domain accuracy remains 99.6%. This suggests that the adult domain requires more training examples to cover its greater variability, likely because VoxCeleb contains a wider range of speakers, ages, accents, and acoustic conditions than the CSLU child speech data.

Although the domain classifier is not perfect, its errors have a smaller impact on speaker verification EER than the raw domain-classification error rate might suggest. This is because misclassification does not necessarily cause a verification error. For example, in some non-target trials, assigning the two utterances to different domains can increase their embedding difference and reduce cosine similarity, which does not harm verification and may even help in some cases.

7.3.2 Adult-Only and Child-Adapted Systems

Table 7.2 reports EER results for adult-domain, child-domain, and age-agnostic systems. The adult-domain systems perform well on VoxCeleb but poorly on younger CSLU age groups. For example, A-SV-Large obtains an EER of 0.80% on VoxCeleb1-O, but its EER is above 30% for the youngest CSLU age groups K00–K02. This confirms the strong mismatch between adult-trained speaker verification models and child speech.

The child-adapted systems substantially improve performance on CSLU. C-SV-Large reduces EER from 30.65% to 8.90% on K00 and from 31.75% to 8.55% on K02 relative to A-SV-Large. However, this improvement comes at the cost of adult-domain degradation. On VoxCeleb1-O, C-SV-Large has an EER of 7.01%, compared with 0.80% for A-SV-Large. Similarly, C-SV-Small has an EER of 10.21% on VoxCeleb1-O, compared with 1.19% for A-SV-Small. These results show that conventional child-domain adaptation improves C-SV but causes substantial source-domain forgetting.

Table 7.2: EER (%) comparison of verification systems. K00–K10 represent increasing child age groups in CSLU. A-SV and C-SV denote adult and children speaker verification systems, respectively. AASV denotes the proposed Age-Agnostic Speaker Verification system. DC denotes the domain classifier, and PreEmbed denotes pretrained speaker embeddings.

#	System	CSLU Children's Speech										VoxCeleb Adult Speech			
		K00	K01	K02	K03	K04	K05	K06	K07	K08	K09	K10	VoxCeleb1-O	VoxCeleb1-E	VoxCeleb1-H
1	A-SV-Large	30.65	30.55	31.75	26.85	23.90	12.55	13.10	9.15	6.15	5.80	6.15	0.80	-	-
2	C-SV-Large	8.90	10.10	8.55	6.40	6.45	3.60	3.20	2.55	2.40	1.60	3.20	7.01	-	-
3	ChildAugment [19]	29.30	29.65	29.95	24.55	23.55	12.30	13.15	8.90	6.10	5.65	5.75	1.19	1.49	2.98
4	AASV-Large	8.90	10.20	8.55	6.40	6.45	3.60	3.20	2.55	2.30	1.70	4.00	1.89	-	-
5	A-SV-Small	31.95	34.40	32.30	29.85	28.80	15.40	14.90	11.60	10.10	9.15	9.60	1.19	1.31	2.48
6	C-SV-Small	9.50	9.50	8.45	6.60	5.60	3.80	3.30	2.80	2.50	1.80	3.00	10.21	9.55	13.38
7	Stable Learning [84]	8.55	9.60	8.40	6.55	6.35	4.00	3.05	2.80	2.15	1.60	3.15	10.26	9.55	13.38
8	WSE [29]	17.25	18.40	16.05	14.30	13.95	7.25	5.80	4.35	3.15	3.55	3.15	1.92	2.13	3.44
9	AASV-Small	9.50	9.60	8.45	6.65	5.55	3.80	3.30	2.80	2.40	1.90	3.80	2.25	-	-
	w/o DC	16.65	17.60	16.35	13.80	13.95	8.00	5.90	4.75	4.45	3.30	4.75	1.85	2.13	3.63
	w/o DC + PreEmbed	19.80	22.15	21.10	19.45	19.65	10.20	9.00	6.00	5.75	4.90	5.10	2.43	2.66	4.56
10	AASV-Small + PreEmbed	11.70	14.45	12.54	11.00	11.45	8.40	7.20	5.00	4.35	3.95	5.15	1.95	-	-

7.3.3 Performance of the Proposed AASV System

The proposed AASV systems provide a better balance between child-domain and adult-domain performance. AASV-Large retains nearly the same performance as C-SV-Large on the CSLU child test sets while reducing the adult-domain degradation. On VoxCeleb1-O, AASV-Large obtains an EER of 1.89%, compared with 7.01% for C-SV-Large. This is a large reduction in source-domain degradation, while child-domain performance remains close to the child-adapted system.

The same trend is observed for the smaller model. AASV-Small maintains performance close to C-SV-Small across the CSLU age groups. For K00–K09, the absolute increase in EER relative to C-SV-Small is within 0.1%, and for K10 it is 0.8%. At the same time, AASV-Small obtains 2.25% EER on VoxCeleb1-O, much lower than the 10.21% EER of C-SV-Small. This shows that the proposed embedding integration strategy substantially reduces catastrophic forgetting compared with direct child-domain adaptation.

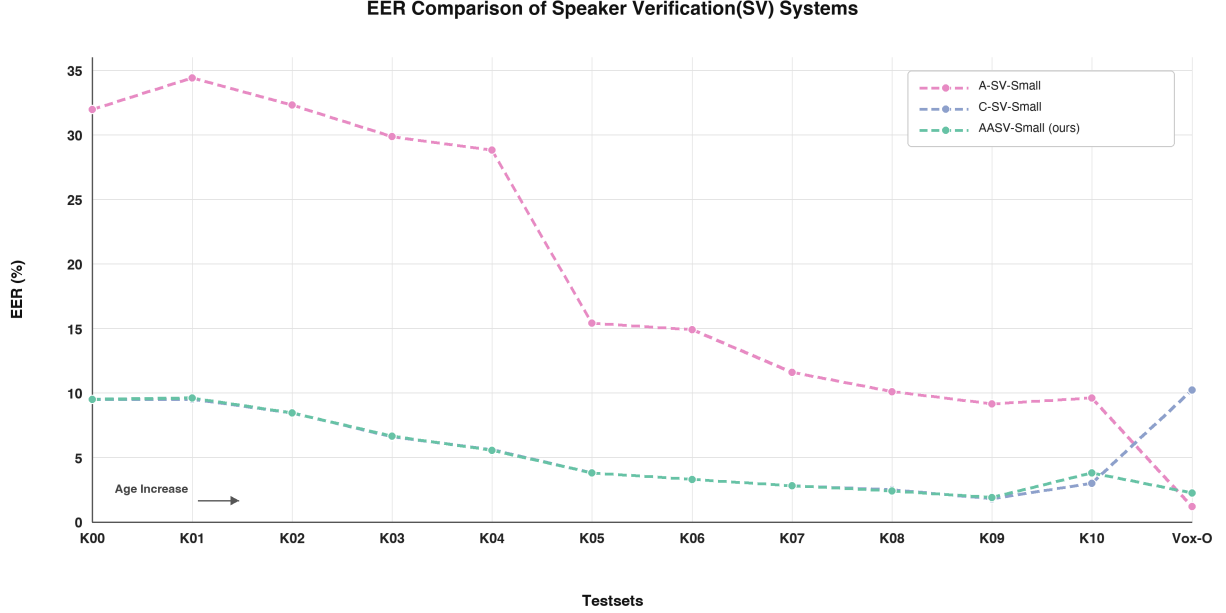


Figure 7.3: Performance of the A-SV-Small, C-SV-Small, and proposed AASV-Small systems across year-wise CSLU test sets and the VoxCeleb1 test set.

Figure 7.3 summarizes the behavior of A-SV-Small, C-SV-Small, and AASV-Small across the child and adult test sets. The adult-trained model performs poorly on younger child age groups but strongly on VoxCeleb. The child-adapted model performs strongly on CSLU but degrades on VoxCeleb. The AASV system follows the child-adapted model closely on CSLU while remaining much closer to the adult-trained model on VoxCeleb. This pattern captures the intended goal of AASV.

Compared with prior methods, the proposed AASV system provides a stronger balance across age groups. ChildAugment [19] preserves adult-domain performance but provides limited improvement on the CSLU child test sets. Stable Learning [84] improves C-SV performance but degrades A-SV performance substantially. WSE [29] provides competitive adult-domain performance, but AASV-Small outperforms WSE on 10 out of 11 CSLU age-group test sets. These comparisons show that age-agnostic integration is a useful alternative to methods that either preserve the source domain at the cost of child-domain performance or improve the child domain at the cost of source-domain retention.

7.3.4 Ablation Study on Embedding Integration

The ablation results in Table 7.2 show that the domain classifier is important for effective age-agnostic integration. When adult- and child-domain embeddings are combined without the domain classifier, performance on the CSLU child test sets drops substantially. For example, the *w/o DC* system has an EER of 16.65% on K00, compared with 9.50% for AASV-Small. Similar degradations appear across most CSLU age groups. This indicates that simple concatenation or unweighted integration is not sufficient. The system needs domain information to decide how much each embedding space should contribute.

The *w/o DC + PreEmbed* and *AASV-Small + PreEmbed* results further show that directly incorporating pretrained adult embeddings during child-domain fine-tuning does not provide the desired balance. Although pretrained embeddings can be useful in some transfer learning settings [85], the results here show that this strategy is less effective than the proposed domain-classifier-guided integration. The AASV system performs best when adult- and child-domain embeddings are kept separate and combined dynamically using domain probabilities.

7.3.5 Discussion

The results in this chapter show that child-domain adaptation creates a clear trade-off: it improves C-SV performance but can substantially degrade A-SV performance. This is the central catastrophic forgetting issue that motivates the final part of the dissertation. However, the results also show that the trade-off can be reduced when adult- and child-domain representations are combined rather than forcing a single adapted model to handle both domains.

The proposed AASV framework works because it preserves two complementary embedding spaces. The adult-domain embedding captures information useful for adult speech, while the child-domain embedding captures information useful for child speech. The domain classifier then provides a soft weighting mechanism that allows the system to emphasize

the appropriate representation for each input. This avoids a hard decision between adult and child systems, and it also avoids fully overwriting the adult-domain model during child adaptation.

This chapter assumes that source-domain information is available when constructing the age-agnostic system. This assumption makes it possible to use adult-domain data along with child-domain data to train the domain classifier. In practice, however, source-domain data may not always be available during adaptation because of privacy, licensing, storage, or access constraints. Also, the AASV framework, maintains two separate model branches for adult and child domain respectively. Hence effectively increasing the number of model parameters. Therefore, in the next chapter we study a more restrictive source-free setting, where the goal is to adapt to a target domain while reducing catastrophic forgetting without access to the original source-domain data and also ensuring minimum increase in the model parameters.

7.4 Summary

This chapter presented an Age-Agnostic Speaker Verification (AASV) system designed to perform robustly across both adult and child speech. The proposed framework combines embeddings from adult-domain and child-domain speaker verification systems using age-domain information estimated by a domain classifier. The resulting representation expands the embedding space and dynamically weights the contribution of adult- and child-domain embeddings.

Experiments on CSLU and VoxCeleb showed that adult-trained systems perform well on adult speech but poorly on child speech, while child-adapted systems improve child speaker verification but substantially degrade adult speaker verification. The proposed AASV systems reduce this trade-off. AASV-Large maintains child-domain performance close to C-SV-Large while reducing VoxCeleb1-O EER from 7.01% to 1.89%. AASV-Small similarly maintains performance close to C-SV-Small on CSLU while reducing VoxCeleb1-O EER from

10.21% to 2.25%.

The main conclusion of this chapter is that AASV can reduce catastrophic forgetting by preserving and integrating both adult- and child-domain representations. This provides a unified model for adult and child speaker verification under the assumption that source-domain information is available. The next chapter removes this assumption and studies source-free adaptation methods for reducing catastrophic forgetting when the original adult source data cannot be accessed.

CHAPTER 8

Cross-Attentive Adapter with Regularized Domain Adaptation for Speaker Verification

8.1 Introduction

The previous chapter addressed age-agnostic speaker verification by combining adult- and child-domain speaker representations. That setting assumes that source-domain information is available when constructing the system. In many practical adaptation scenarios, however, the original source-domain data may not be accessible. Large-scale adult speaker verification datasets may be unavailable during adaptation because of storage constraints, licensing restrictions, privacy concerns, or deployment-time limitations. In such cases, we may only have access to a pretrained adult speaker verification model and target-domain data, but not to the adult speech data originally used to train the model.

Catastrophic forgetting refers to the degradation of previously learned knowledge when a model is adapted to a new task or domain [32, 33]. In the speaker verification setting considered here, adaptation to a target domain can improve target-domain equal error rate (EER), but may substantially degrade EER on the original adult source domain. This creates the stability–plasticity tension: the adapted model should be plastic enough to learn target-domain speaker-discriminative information, but stable enough to retain the source-domain information learned by the pretrained model.

The main problem addressed in this chapter is therefore: *how can a speaker verification model be adapted to a target domain without access to source-domain training data, while reducing catastrophic forgetting on the source domain?* This problem is especially important

for adult-to-child speaker verification adaptation because of the significant mismatch between adult and child speech. A source-free method must improve performance on child speech while not letting adult speech performance degrade, without relying on adult source examples during adaptation.

To address this problem, this chapter proposes a source-free adaptation framework that uses pretrained-model embeddings (PMEs) extracted from target-domain inputs. The central hypothesis is that when target-domain utterances are passed through the frozen source-domain model, the resulting embeddings preserve useful information about the source model’s representation space. These embeddings are not substitutes for the original source data, but they provide a proxy for the pretrained model’s source-domain knowledge.

This chapter proposes three strategies: (i) **Cross-Attentive (CA) Adapter**, a hybrid of knowledge distillation and adapter methods; (ii) **K-means-based modified Elastic Weight Consolidation (K-EWC)**, for parameter regularization, and (iii) a **Moment-Matching Regularization Loss**, which constrains fine-tuning by aligning learnable target-domain embeddings distribution with PME distributions.

8.2 Methods

8.2.1 Source-Free Adaptation Setting

Let the pretrained source-domain speaker verification model be denoted by f_{θ^*} , where θ^* represents the parameters learned from adult source-domain training. In the source-free adaptation setting, the original source-domain training data are unavailable. The adaptation procedure has access only to the pretrained model f_{θ^*} and target-domain data D_{Tar} . The goal is to obtain an adapted model f_{θ} that performs well on the target-domain speaker verification task while retaining performance on the source-domain speaker verification task.

The target-domain data are used to optimize the model for speaker identification during adaptation. As described in Chapter 3, speaker embeddings are then evaluated using speaker verification trials and cosine similarity scoring. Target-domain EER measures adaptation

performance, while source-domain EER measures how much source-domain knowledge is retained. A successful source-free adaptation method should therefore reduce target-domain EER without causing a large increase in source-domain EER.

The proposed methods rely on pretrained-model embeddings (PMEs). For each target-domain utterance $x \in D_{Tar}$, the frozen pretrained source model produces an embedding

$$e_T = f_{\theta^*}(x). \quad (8.1)$$

Because PMEs are produced by the unadapted source-domain model, PMEs reflect the source representation space, i.e., how the source model represents target-domain inputs before adaptation. The methods in this chapter use these PMEs to anchor adaptation by preserving the knowledge of that source representation space.

8.2.2 Overview of the Proposed Framework

Figure 8.1 shows the proposed source-free adaptation framework. The framework contains two connected ideas. First, the Cross-Attentive (CA) adapter uses both a frozen teacher branch and a learnable student branch. The teacher branch provides fixed source-model representations, while the student branch adapts to the target domain. The CA adapter dynamically combines information from both branches during pooling. Second, K-means-based Elastic Weight Consolidation (K-EWC) and moment-matching (MM) regularization use PMEs to constrain adaptation and reduce forgetting.

The key design choice is that the original source-domain data are never required during adaptation. The source-domain information used by the method is obtained only through the pretrained model’s response to target-domain inputs. This makes the framework suitable for scenarios where source data cannot be stored, redistributed, or accessed during adaptation.

8.2.3 Cross-Attentive (CA) Adapter

The CA adapter, as seen in Figure 8.1 (a), modifies the attentive statistics pooling stage of the ECAPA-TDNN [2] speaker embedding network. The model is organized as a Siamese-style

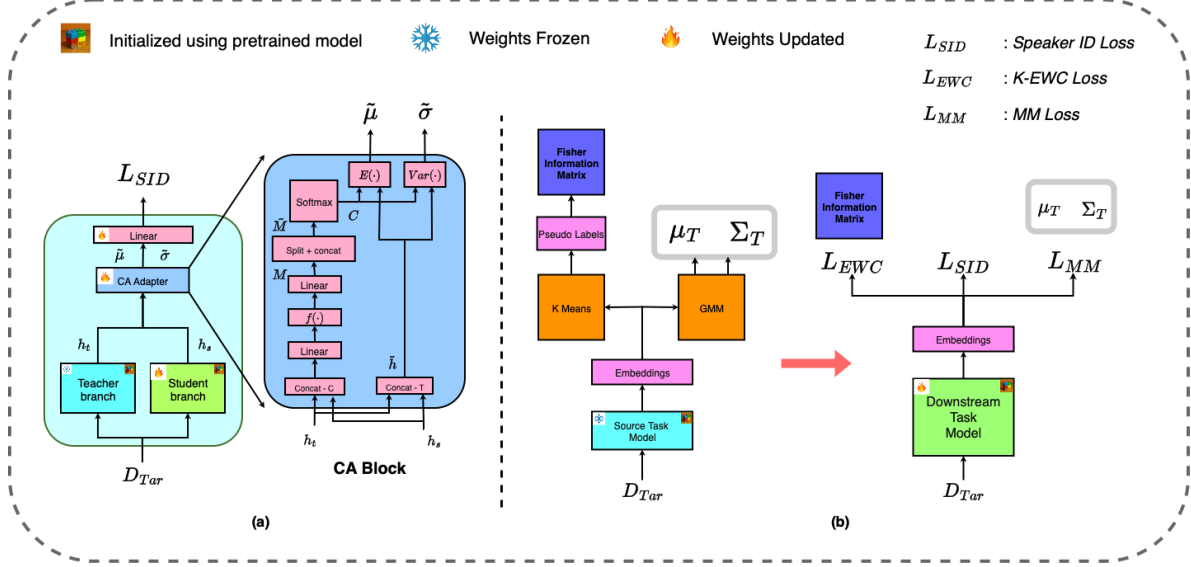


Figure 8.1: Overview of the proposed approaches. The CA adapter framework is shown in (a). Hidden representations h_t and h_s have dimension $C \times T$. In the CA Block, “Concat-C” and “Concat-T” refer to concatenation operation on the hidden representations along the C (channel) and T (time) axes, respectively. Estimation of Fisher Information Matrix (FIM) for K-EWC using pseudo-labels / Estimation of pseudo μ_T and Σ_T for MM regularization is shown in (b). D_{Tar} refers to the target domain data.

framework with a frozen teacher branch and a learnable student branch. Both branches are initialized from the pretrained source-domain model. During adaptation, the teacher branch remains fixed, while the student branch is updated using target-domain data.

Let h_t and h_s denote the hidden representations from the teacher and student branches, respectively. In this dissertation, these representations are taken from the portion of the ECAPA-TDNN network up to the multi-layer feature aggregation module. Each representation has dimension $1536 \times T$, where T is the number of frames. The purpose of the CA adapter is to compute attentive statistics over the combined teacher (h_t) and student (h_s) representations so that the adapted embedding can use both fixed source-model information and learnable target-domain information.

The hidden representations from the teacher and student branches - h_t , h_s , each $\in$

$\mathbb{R}^{1536 \times T}$, are stacked channel-wise (“Concat-C” within **CA Block** in Figure. 8.1 (a)), followed by a linear transformation with parameters $W \in \mathbb{R}^{l \times 3072}$ and $b \in \mathbb{R}^{l \times 1}$ to down-project into a smaller latent space with dimension l , using a non-linearity $f(\cdot)$. After the non-linearity $f(\cdot)$, the self-attention scores M across both hidden representations are calculated through a second linear transformation with parameters $V \in \mathbb{R}^{3072 \times l}$ and $k \in \mathbb{R}^{3072}$.

$$M = V \left[f \left(W \begin{bmatrix} h_t \\ h_s \end{bmatrix} + b \right) \right] + k, \quad M \in \mathbb{R}^{3072 \times T}. \quad (8.2)$$

The self-attention scores M is reshaped into $\tilde{M}$ to ensure that channel information across the corresponding teacher and student representations are aligned along the channel dimension.

$$\tilde{M} = \begin{bmatrix} M_{1:1536,1:T} & M_{1537:3072,1:T} \end{bmatrix}, \quad \tilde{M} \in \mathbb{R}^{1536 \times 2T}. \quad (8.3)$$

A channel-wise softmax is then applied over time to obtain normalized attention weights:

$$C = \text{softmax}(\tilde{M}), \quad C \in \mathbb{R}^{1536 \times 2T}. \quad (8.4)$$

The original teacher and student hidden representations are also concatenated along the time dimension (“Concat-T” within **CA Block** in Figure. 8.1 (a)):

$$\tilde{h} = \begin{bmatrix} h_t & h_s \end{bmatrix}, \quad \tilde{h} \in \mathbb{R}^{1536 \times 2T}. \quad (8.5)$$

The normalized attention scores C , which represent the importance of each frame given the channel, are used to compute mean ($\tilde{\mu}$) and standard deviation ($\tilde{\sigma}$).

$$\tilde{\mu} = \sum_{t=1}^{2T} C \odot \tilde{h}, \quad \tilde{\mu} \in \mathbb{R}^{1536}, \quad (8.6)$$

$$\tilde{\sigma} = \sqrt{\sum_{t=1}^{2T} C \odot \tilde{h}^2 - \tilde{\mu}^2}, \quad \tilde{\sigma} \in \mathbb{R}^{1536}. \quad (8.7)$$

The resulting mean and standard deviation vectors are concatenated and passed through a linear layer to produce a 192-dimensional speaker embedding.

Unlike the age-domain classifier based approach proposed in Chapter 7 (Figure 7.1), which requires source data to train the domain classifier, the CA adapter can be viewed as

a source-free mechanism for balancing retention and adaptation. The frozen teacher branch preserves the pretrained model’s representation, while the learnable student branch adapts to the target domain. The cross-attentive pooling mechanism allows the final embedding to use information from both branches instead of discarding the source representation during adaptation. In the implementation used here, the CA block increases the ECAPA-TDNN model size from approximately 20.8M parameters to 22.3M parameters, unlike 2x increase in the case of age-domain classifier based approach proposed in Chapter 7 .

8.2.4 K-Means-Based Elastic Weight Consolidation

Elastic Weight Consolidation (EWC) is a regularization method designed to reduce catastrophic forgetting by penalizing changes to parameters that are important for the source task [40]. EWC was selected as the starting point regularization in this dissertation because it directly addresses the problem of catastrophic forgetting during model adaptation. EWC does not require source-domain examples to be replayed during adaptation. Instead, it constrains changes to model parameters according to their estimated importance, conventionally represented by the diagonal of the Fisher information matrix (FIM) [86].

Let θ_i denote the i th parameter of the adapted model and let θ_i^* denote the corresponding pretrained source-model parameter. EWC adds the following regularization term during adaptation:

$$\mathcal{L}_{\text{EWC}} = \sum_i F_i (\theta_i - \theta_i^*)^2, \quad (8.8)$$

where F_i is the diagonal element of the FIM corresponding to parameter θ_i . The FIM elements F_i are a measure of how much the log-likelihood of the data changes when a given i^{th} parameter is perturbed. More specifically, larger values of F_i indicate that the corresponding parameter is more important for the source task, so changing that parameter is penalized more strongly.

However, standard EWC requires source-domain data to estimate the FIM. This is not possible in the source-free setting considered in this chapter. This limitation motivated the investigation of whether an EWC-inspired parameter-importance estimate could instead be

constructed from the target data. K-means clustering was therefore used to obtain pseudo-labels from the PME, i.e., target-domain inputs representations from the pretrained model, enabling the estimation of parameter importance, i.e., FIM, without access to source data examples. Once the approximate FIM is obtained, the regularization term in Eq. 8.8 is used during target-domain adaptation. K-EWC therefore preserves the central idea of EWC, namely constraining parameters that are important for the source model, while removing the requirement that source-domain samples be available during adaptation.

8.2.5 Moment-Matching Regularization

The second regularization strategy is moment matching (MM). The goal of MM regularization is to prevent the learned target-domain embeddings from drifting too far from the pretrained model’s embedding distribution. If the adapted embeddings move to a very different region of the representation space, the model may improve on the target domain but forget the source-domain structure learned by the pretrained model.

As with K-EWC, MM regularization uses PMEs as a proxy for source-domain knowledge. The PMEs are used to fit a Gaussian Mixture Model (GMM), with the number of mixture components selected using the Akaike Information Criterion (AIC). After fitting the GMM, its overall mean vector and covariance matrix by combining the component-specific statistics according to their mixture weights. These estimates of the mean vector μ_T and covariance matrix Σ_T are used as the approximation for the pseudo-source distribution. During adaptation, the learned student embeddings within a batch have mean μ_S and covariance Σ_S . The MM regularization loss is defined as

$$\mathcal{L}_{\text{MM}} = \|\mu_T - \mu_S\|_2^2 + \|\Sigma_T - \Sigma_S\|_F^2 \quad (8.9)$$

where $\|\cdot\|_2$ denotes the L_2 norm and $\|\cdot\|_F$ denotes the Frobenius norm. This loss encourages the adapted embeddings to remain close to the pseudo-source distribution while still allowing the model to learn target-domain speaker-discriminative information.

8.2.6 Overall Training Objective

As shown in Figure 8.1 (b), the overall adaptation objective combines the speaker identification loss with the two source-free regularization losses:

$$\mathcal{L} = \mathcal{L}_{\text{SID}} + \lambda_{\text{EWC}}\mathcal{L}_{\text{EWC}} + \lambda_{\text{MM}}\mathcal{L}_{\text{MM}}. \quad (8.10)$$

Here, $\mathcal{L}_{\text{SID}}$ is the speaker identification loss on the target-domain data, while λ_{EWC} and λ_{MM} control the strength of the K-EWC and MM regularization terms. The methods are evaluated both individually and in combination with the CA adapter.

As described in Chapters 2 and 3 - the experiments use 80-dimensional filterbank features extracted using a 25 ms frame length and 10 ms hop size. The underlying speaker embedding model is ECAPA-TDNN [2], implemented using SpeechBrain [78]. The pretrained source model is trained on VoxCeleb 1 and 2 [51, 52], and target-domain adaptation is performed on MyST [50] and CNCeleb [53]. MyST represents a cross-age adaptation scenario, while CNCeleb represents a cross-lingual adaptation scenario. The source-domain performance after adaptation is evaluated using VoxCeleb-O evaluation set in order to quantify catastrophic forgetting.

8.3 Results and Discussion

8.3.1 Evaluation Overview

The experiments evaluate each adaptation method using target-domain EER and source-domain EER. Table 8.1 summarizes the main results. Here, results are reported as Target/Source EER. For example, in Table 8.1, in the MyST/VoxCeleb-O setting, the first number reports EER on the MyST evaluation set and the second number reports EER on VoxCeleb-O. Lower target-domain EER indicates better adaptation, while lower source-domain EER indicates less catastrophic forgetting. As discussed in Chapter 2 - Sections 2.3.3 and 2.3.5, models are adapted using four subsets each of MyST (MyST-1 to -4) and CNCeleb (CNCeleb-1 to -4). The amount of adaptation data increases from subsets -1 to -4.

The no-adaptation VoxCeleb-pretrained model provides a useful reference point: it obtains 0.89% EER on VoxCeleb-O, 17.48% EER on MyST, and 15.26% EER on CNCeleb. These values show that the pretrained adult English model performs very well on the source-domain evaluation set but is less effective on the target-domain tasks. Therefore, the purpose of adaptation is to reduce target-domain EER while keeping source-domain EER as close as possible to the no-adaptation source performance.

In Table 8.1, the **Baseline** system is trained only on target-domain data, while **Finetune** initializes from the VoxCeleb-pretrained model and adapts using target-domain speaker identification loss. The rows labeled **+MM** and **+K-EWC** add the corresponding source-free regularization term to **Finetune**. **CA** denotes the proposed Cross-Attentive adapter, and the remaining rows combine **CA** with **MM** and/or **K-EWC**.

Table 8.1: Model performance in terms of Equal Error Rate (%EER). Results are reported as Target/Source %EER. For MyST/VoxCeleb-O, the first number is MyST EER and the second number is VoxCeleb-O EER. For CNCeleb/VoxCeleb-O, the first number is CNCeleb EER and the second number is VoxCeleb-O EER. The target training set duration increases from Split-1 to Split-4.

Method	MyST/VoxCeleb-O				CNCeleb/VoxCeleb-O			
	MyST-1	MyST-2	MyST-3	MyST-4	CNCeleb-1	CNCeleb-2	CNCeleb-3	CNCeleb-4
1 Baseline	21.84/28.97	12.72/18.08	10.07/21.10	8.64/21.18	32.06/25.16	19.51/15.77	17.13/13.23	16.17/11.76
2 G-IFT	14.22/8.00	7.03/8.94	5.49/12.51	5.42/11.23	-	-	-	-
3 Sparse FilterBank	-	-	-	-	-	-	-	12.25/10.81
4 Finetune	20.30/7.73	7.22/7.60	5.67/8.39	5.57 /14.91	17.96/6.05	10.00/3.52	10.57/4.63	10.98/6.52
5 +MM	22.58/6.92	7.64/7.06	5.54* /11.14	5.59/15.57	18.60/5.33	10.19/3.56	10.89/5.23	10.94/6.16
6 +K-EWC	21.00/5.49	8.86/6.00	7.41/7.18	7.26/7.76	14.61* / 2.04*	10.40/1.90	10.20/2.63	10.26/2.85
7 CA	16.53* /4.16	6.93* /5.24	5.56/6.55	5.78/11.15	16.99/4.01	10.14/3.22	10.36/3.96	10.56/3.97
8 +MM	16.79/4.42	7.04/5.51	5.75/6.56	5.70/10.05	16.96/4.17	10.16/3.24	10.50/4.14	10.44/3.94
9 +K-EWC	17.29/3.89	7.64/3.11	6.54/3.83	6.28/ 4.39*	15.47/3.02	9.96/ 1.80*	10.14* / 2.45*	10.27/2.73
10 +K-EWC+MM	17.29/ 3.88*	7.58/ 3.01*	6.57/ 3.74*	6.37/4.40	15.46/3.03	9.95* /1.81	10.19/2.52	10.17* / 2.72*

8.3.2 Baseline and Fine-Tuning Behavior

The Baseline system, trained only on target-domain data, performs poorly on the source-domain VoxCeleb-O evaluation set. This is expected because the model is not trained to preserve adult English source-domain knowledge. The effect is especially clear for MyST, where source-domain EER is 28.97% for MyST-1 and remains above 18% for larger splits. Similarly in the case of CNCeleb, the Baseline model performance on VoxCeleb-O evaluation set remains above 11.76%.

Fine-tuning improves target-domain performance compared with the no-adaptation pre-trained model, across all adaptation splits, except for “MyST-1” (no-adaptation value of 17.48% to 20.30%) and “CNCeleb-1” (no adaptation value of 15.26% to 17.96%). In the rest of the MyST settings, fine-tuning reduces target-domain EER from the no-adaptation value of 17.48% to 7.22% on MyST-2, 5.67% on MyST-3, and 5.57% on MyST-4. However, this improvement comes with a source-domain cost. In MyST-4, the VoxCeleb-O EER increases to 14.91% (from the no adaptation EER of 0.89%), showing substantial catastrophic forgetting. Similar pattern is observed in the case of CNCeleb, with CNCeleb-4 resulting in an EER of 10.98% on CNCeleb evaluation set, while the VoxCeleb-O EER degrading to 6.52%. This confirms that target-domain improvement alone is not sufficient for evaluating adaptation.

The relationship between target-domain training data size and source-domain degradation is different for MyST and CNCeleb. In the MyST experiments, larger target-domain adaptation sets generally create stronger drift away from the source-domain representation, especially in the high-resource MyST-4 condition. This reflects the large mismatch between child speech and adult English speech. In the CNCeleb experiments, source-domain degradation is less severe because both CNCeleb and VoxCeleb contain adult speech. The cross-lingual mismatch is still important, but the acoustic mismatch is smaller than in the adult-to-child setting.

8.3.3 Effect of Source-Free Regularization

The MM and K-EWC results show two different ways of using PME to reduce forgetting. MM regularization constrains the learned target-domain embedding distribution to remain close to the PME distribution. This helps source-domain retention in some low-resource cases, but its behavior is not uniform across all splits. In MyST-1 and MyST-2, MM reduces source-domain EER compared with Finetune. However, for larger MyST splits, MM does not consistently prevent source-domain degradation. This suggests that moment matching is sensitive to the quality of the pseudo-source distribution estimate and to the dynamics of the adaptation schedule.

K-EWC provides more consistent source-domain retention. Across both MyST and CNCeleb, adding K-EWC to Finetune reduces VoxCeleb-O EER relative to standard finetuning. For example, in the MyST-4 condition, K-EWC reduces source-domain EER from 14.91% to 7.76%. In the CNCeleb conditions, K-EWC also substantially reduces source-domain EER, reaching 2.04% on CNCeleb-1 and 1.90% on CNCeleb-2. These results suggest that estimating parameter importance using K-means pseudo labels derived from PMEs is an effective source-free approximation to the role played by source data in standard EWC.

The trade-off is that K-EWC can reduce target-domain performance. This is expected because stronger source-domain retention restricts how much the model can move toward the target domain. The results therefore illustrate the core stability-plasticity trade-off: methods that strongly preserve the source model may limit target-domain adaptation, while methods that adapt aggressively may forget the source domain.

8.3.4 Effect of the Cross-Attentive Adapter

The CA adapter provides a different mechanism for reducing forgetting. Rather than only penalizing parameter changes or embedding distribution drift, it explicitly combines information from a frozen teacher branch and a learnable student branch. This allows the adapted representation to retain information from the pretrained model while still learning target-

domain structure.

For MyST, the CA adapter improves target-domain EER compared with Finetune in MyST-1, MyST-2, and MyST-3. It also reduces source-domain degradation compared with standard fine-tuning in those cases. The improvement is particularly important in the low-resource MyST-1 setting, where CA reduces target-domain EER from 20.30% to 16.53% and source-domain EER from 7.73% to 4.16%. This shows that cross-attentive use of teacher and student representations is especially useful when target-domain data are limited.

In the high-resource MyST-4 condition, CA obtains strong target-domain performance but still shows source-domain degradation. This indicates that the CA adapter alone is not always sufficient to preserve the source domain when the target-domain adaptation pressure is strong (i.e., more target domain data). In the MyST-4 setting, adding K-EWC to CA reduces source-domain EER substantially, from 11.15% with CA alone to 4.39%, although this comes with some target-domain degradation. The combination of CA with K-EWC and MM gives the strongest source-domain retention for MyST-1 through MyST-3, with VoxCeleb-O EERs of 3.88%, 3.01%, and 3.74%, respectively.

For CNCeleb, the CA adapter generally reduces source-domain degradation while maintaining competitive target-domain performance. Because CNCeleb and VoxCeleb are both adult speech datasets, the source-target mismatch is smaller than in the MyST case. This makes it easier to find methods that improve target-domain performance without severely degrading source-domain performance. In this setting, CA combined with K-EWC or with both K-EWC and MM achieves strong source-domain retention and competitive target-domain EER.

8.3.5 Cross-Age Versus Cross-Lingual Adaptation

A comparison of MyST and CNCeleb highlights an important distinction between cross-age and cross-lingual adaptation. In the CNCeleb experiments, several methods improve target-domain performance while maintaining relatively low VoxCeleb-O EER. This suggests that adapting between adult English and adult Chinese speech produces a manageable domain

shift for the source-free methods studied here.

In contrast, the MyST experiments show a stronger trade-off between target-domain improvement and source-domain retention. Methods that achieve the best MyST EER do not always achieve the best VoxCeleb-O EER. This is consistent with the broader argument of the dissertation: adult-to-child adaptation is not simply another domain adaptation problem. The developmental differences between adult and child speech create a stronger mismatch than many adult-to-adult domain shifts. Therefore, preserving adult-domain performance while adapting to child speech requires explicit mechanisms to control forgetting.

These results also clarify the relationship between this chapter and the previous chapter. Chapter 7 addressed adult-child robustness by combining source- and target-domain systems under access to source-domain information. This chapter removes that assumption and shows that source-free adaptation is possible by using PME as a proxy for source-domain knowledge. Although the source-free setting is more restrictive, the proposed CA, K-EWC, and MM approaches substantially reduce source-domain degradation compared with conventional target-domain fine-tuning.

8.4 Summary

This chapter addressed catastrophic forgetting in speaker verification under source-free domain adaptation. The setting assumes that a pretrained adult source-domain model is available, but the original source-domain training data are not available during adaptation. This constraint is important for practical speaker verification systems, where source data may be restricted by privacy, licensing, storage, or deployment considerations.

The chapter proposed three source-free strategies. First, the Cross-Attentive adapter combines a frozen teacher branch with a learnable student branch, allowing the adapted model to balance source-model representations and target-domain learning. Second, K-means-based EWC estimates source-parameter importance using pseudo labels obtained by clustering pretrained-model embeddings. Third, moment-matching regularization constrains

the adapted embedding distribution to remain close to a pseudo-source distribution estimated from pretrained-model embeddings.

The experiments showed that conventional fine-tuning improves target-domain performance but can substantially degrade source-domain performance, especially in adult-to-child adaptation. K-EWC consistently reduced source-domain forgetting across both cross-age and cross-lingual target domains. The CA adapter improved target-domain performance and mitigated forgetting in several settings, particularly under low-resource target-domain adaptation. Combining CA with K-EWC and MM further improved source-domain retention in many cases.

Overall, this chapter shows that source-free adaptation can reduce catastrophic forgetting in speaker verification without requiring access to source-domain data. Together with the preceding chapters, these results complete the dissertation’s progression from characterizing developmental variability in child speech to building robust, efficient, age-agnostic, and source-free speaker verification systems.

CHAPTER 9

Summary and Future Work

This dissertation addressed the problem of robust speaker verification for children. The central motivation was that children are not simply smaller versions of adult speakers. Their speech differs from adult speech because of ongoing anatomical, acoustic, and articulatory development, and their voices continue to change over time. At the same time, child speech data is limited compared with large-scale adult speech data. These factors pose challenges towards building efficient speech technology systems for children.

Among speech various technology applications, with speaker verification as the focus, in this dissertation we followed a progression from analysis to adaptation. First, we quantified developmental variability in children’s speech using longitudinal acoustic and articulatory measurements. Second, we studied age-related robustness in children speaker verification, especially when enrollment and test utterances are separated by one or more years. Third, we investigated efficient adult-to-child adaptation, where adult-trained speaker verification models are adapted to the low-resource child speech domain. Fourth, we addressed the need for unified systems that perform well on both adult and child speech. Finally, we studied source-free adaptation methods that reduce catastrophic forgetting when the original adult source-domain data are unavailable during adaptation.

Together, these studies provide a unified view of child speaker verification as a developmental, low-resource, and domain-adaptation problem. The results show that improving child speaker verification requires more than collecting child speech data or fine-tuning adult-trained models. It requires modeling the specific forms of variability present in children’s speech, designing adaptation strategies that use limited target-domain data efficiently, and

controlling the loss of source-domain knowledge during adaptation.

9.1 Major Findings and Contributions

9.1.1 Quantifying Developmental Variability in Child Speech

The first contribution of this dissertation was a longitudinal analysis of acoustic and articulatory development in children, presented in Chapter 4. This analysis examined children recorded yearly from Grade 1 to Grade 4 and focused on four corner vowels. The analysis showed that children’s speech development contains both systematic trends and substantial individual variability. This contribution is important because it establishes the developmental basis for the rest of the dissertation. A robust children speaker verification system (and any speech technology application in general) must therefore account for the fact that a child’s voice can change meaningfully across school years.

9.1.2 Improving Age-Related Robustness in Children Speaker Verification

The second contribution of this dissertation was the development and evaluation of methods for age-related robustness in children speaker verification, presented in Chapter 5. A longitudinal evaluation setup was introduced to measure both intra-year and inter-year verification performance. This evaluation framework made it possible to directly test whether a system can verify the identity of a particular child when enrollment and test speech are collected from different grades. The results showed that inter-year verification is substantially more difficult than intra-year verification. Conventional fine-tuning improved performance in some same-year conditions but did not consistently improve robustness across years. This finding is important because it shows that child-domain adaptation and age-robust adaptation are not identical objectives. A model can become better matched to child speech in general while still remaining sensitive to developmental changes within the same child.

9.1.3 Efficient Adult-to-Child Adaptation Under Limited Child Speech Data

The third contribution was an efficient adaptation framework for transferring adult speaker verification knowledge to child speech, presented in Chapter 6. The motivation was that adult speech datasets are much larger than child speech datasets, so adult-trained models provide a strong starting point. However, because child speech data is limited, conventional fine-tuning may not use the available target-domain data efficiently and may overfit to the available child speakers. To address this problem, Chapter 6 proposed an architecture-agnostic (ECAPA-TDNN, ResNet, and X-vector) framework using the Gated Linear Unit adapter with Iterative Fine-Tuning (G-IFT). The GLU adapter provides a gated mechanism that controls how information from the pretrained adult speaker embedding model is passed forward during child-domain adaptation. The iterative fine-tuning strategy separates updates to the classifier, adapter, and pretrained embedding model, allowing the system to adapt in a more structured way.

9.1.4 Building Unified Age-Agnostic Speaker Verification Systems

The fourth contribution was the development of an Age-Agnostic Speaker Verification (AASV) framework, presented in Chapter 7. Simple adaptation of adult speech trained systems to child speech improves child speaker verification performance. However, such adaptation degrades performance on the original adult domain. Hence, the child-adapted system may no longer function well as a general speaker verification system for both adults and children. Chapter 7 addressed this problem by training an age domain classifier, using adult speech and child speech data. The domain classifier was used to estimate age-domain information which was then used to dynamically weight and concatenate adult- and child-domain speaker representations. The resulting speaker representation preserved information from both domains. Hence, the proposed AASV system maintained a strong child-domain performance while recovering much of the adult-domain performance lost through adaptation.

9.1.5 Addressing Catastrophic Forgetting in Source-Free Adaptation

The final contribution of this dissertation was a source-free domain adaptation framework for reducing catastrophic forgetting in speaker verification, presented in Chapter 8. While Chapter 7 assumed access to source-domain information when constructing a unified age agnostic system, this assumption may not hold in practical deployment. Large-scale adult source data may be unavailable during adaptation because of privacy restrictions, licensing constraints, storage limitations, or deployment-time access constraints. Chapter 8 therefore considered the more restrictive setting where the pretrained adult speaker verification model is available, but the original adult source-domain data are not. The central idea was to use pretrained-model embeddings (PMEs) extracted from frozen source model using target-domain inputs as a proxy for source-domain knowledge. Three source-free strategies were studied: a Cross-Attentive (CA) adapter, K-means-based Elastic Weight Consolidation (K-EWC), and moment-matching (MM) regularization. Across two adaptation tasks (Adult English $\rightarrow$ Child English, and Adult English $\rightarrow$ Adult Chinese), these methods reduced source-domain degradation compared with conventional fine-tuning, showing that catastrophic forgetting can be mitigated even without access to source-domain data.

9.2 Remaining Challenges

Although the dissertation makes progress toward robust child and age-agnostic speaker verification, several challenges remain.

First, the longitudinal acoustic and articulatory analysis was conducted on a small number of children and a limited set of words. This allowed detailed within-child analysis, but larger longitudinal corpora are needed to determine how broadly the observed developmental patterns generalize. The analysis also focused on selected acoustic and tongue-shape measures. Analysis on a wide range of acoustic contexts (different vowels, consonants, etc) and articulatory measures would be beneficial.

Second, the longitudinal child speaker verification evaluation set introduced in this disser-

tation is valuable for studying age-related robustness, but it remains limited in scale. Inter-year child speaker verification would benefit from larger datasets containing more speakers, more recording sessions, longer time spans, and a wider range of speaking styles. Such data would make it possible to study whether the same conclusions hold under more realistic deployment conditions.

Third, the proposed age-robustness and adaptation methods were evaluated on specific datasets and model families. Although G-IFT was tested across ECAPA-TDNN, ResNet, and X-vector systems, other modern speaker verification architectures and self-supervised speech representations should also be studied. Similarly, the source-free adaptation methods in Chapter 8 were developed around ECAPA-TDNN, and their generality across other speaker embedding frameworks remains an open question.

9.3 Future Work

The results of this dissertation suggest several future research directions.

9.3.1 Larger Longitudinal Acoustic and Articulatory Studies

A natural extension of the developmental analysis is to study larger longitudinal child speech corpora. Future work should include more children, more words, more phonetic contexts, and a broader range of speaking styles. Systematic variation of phonetic context would make it possible to separate developmental effects from context-dependent coarticulatory effects. In addition, combining ultrasound data with other imaging modalities, such as MRI, could provide estimates of vocal tract area functions and a more complete view of the relationship between vocal tract growth, articulation, and acoustics.

Future work should also explore richer articulatory representations. TCD and TCP provide compact and interpretable summaries of tongue shape, but they do not capture all aspects of articulation. Incorporating additional articulatory measures, including tongue contour dynamics, jaw movement, lip posture, and vocal tract geometry, could improve

understanding of how children achieve stable acoustic targets during development. Such analysis would also provide stronger motivation for speech models that are robust to developmental changes.

9.3.2 More Realistic Longitudinal Speaker Verification Evaluation

The inter-year evaluation framework introduced in this dissertation provides a useful first step for age-robust child speaker verification, but future work should expand it. Larger longitudinal speaker verification datasets are needed to study verification across multiple years, recording conditions, devices, tasks, and speaking styles. Such datasets would allow researchers to evaluate whether age-robust systems generalize beyond controlled speech tasks and small speaker sets.

9.3.3 Improved Age-Robust and Age-Conditioned Representations

The results in Chapter 5 show that age-related robustness requires explicit attention to developmental variability. Future work should explore representations that more directly disentangle speaker identity from age-related acoustic change. This could include age-conditioned embedding models, adversarial approaches that suppress age information while preserving speaker identity, or contrastive objectives that use longitudinal recordings of the same child as positive pairs across time.

Synthetic augmentation is another promising direction. The SAA (from Chapter 5) results suggest that synthetic speech can improve inter-year robustness, but it can also introduce a trade-off with intra-year performance. Future work should study synthesis methods that more carefully control the relationship between speaker identity and age-related variation. Ideally, augmentation methods should vary age-sensitive characteristics while preserving speaker identity in a controlled and interpretable way.

9.3.4 Refining Efficient Adaptation Methods

The G-IFT framework showed that staged adaptation can improve low-resource adult-to-child transfer. Future work should further investigate the relative strengths of G-IFT-1 and G-IFT-2. The results suggest that more separated update schedules may be useful in very low-resource settings, while more joint updates may be sufficient when more target-domain data are available. A future system could choose the adaptation schedule automatically based on the amount and diversity of available child speech.

The framework should also be extended to other low-resource speaker verification settings. Examples include speech from individuals with speech disorders, accented speech, elderly speech, and other populations for which large-scale speaker verification training data are limited. Such studies would test whether the principles developed for child speech adaptation generalize to other underrepresented speech domains.

9.3.5 Toward Stronger Unified Age-Agnostic Systems

The age-agnostic systems in this dissertation show that adult- and child-domain representations can be combined to reduce the trade-off between A-SV and C-SV. Future work should focus on more effective extraction of age-domain information from speech, for example, move from binary adult-child domain information toward continuous or multi-stage age representations. Children’s speech changes gradually, and older children’s speech may be closer to adult speech than younger children’s speech. Hence, a more fine-grained age representation could allow the model to adjust its embedding integration more smoothly across developmental stages. Another direction is to build a single unified model that learns shared and age-specific speaker factors jointly, rather than combining separately trained adult and child systems after the fact.

9.3.6 Generalizing Source-Free Catastrophic Forgetting Methods

The source-free methods in Chapter 8 reduce catastrophic forgetting without access to source-domain data. Future work should extend these methods beyond ECAPA-TDNN to evaluate their generality across a wider range of speaker verification frameworks, including ResNet-based systems, X-vector systems, and self-supervised speech models. This is important because practical speaker verification systems may use different embedding architectures depending on deployment constraints. Future work should also examine other ways of approximating source-domain knowledge. PME provides one useful proxy, but a stronger understanding of when PMEs preserve useful source information, and when they fail, would help make source-free adaptation more reliable.

Memory Aware Synapses (MAS) [44] is a closely related parameter regularization approach that estimates the importance of each model parameter according to the sensitivity of the learned output function to changes in that parameter. In contrast to loss-based importance estimation methods, MAS does not use ground-truth class labels or the task loss, and its parameter-importance values can be accumulated online as samples are observed. When a new task is learned, changes to parameters considered important for previously encountered tasks are penalized. The importance values estimated across successive tasks can then be accumulated to preserve previously acquired knowledge. Although MAS does not require class labels or access to the original source-training examples, its parameter importance values must either be computed using a set of input samples or have been computed and retained previously. Because target-domain speech is available during adaptation, MAS could potentially be applied in the source-free setting by estimating the importance of the source model’s parameters with respect to its behavior on the target-domain samples. This dissertation did not investigate this direction. Future work could evaluate such a target-conditioned MAS approach and investigate whether PMEs could be incorporated into its importance estimation procedure to preserve source-domain speaker-discriminative information more effectively.

Finally, source-free adaptation should be studied in continual and lifelong learning set-

tings. In real deployments, a speaker verification system may need to adapt repeatedly to new users, new age groups, new languages, and new recording conditions. Each adaptation step introduces the possibility of forgetting previous domains. Extending the methods in this dissertation to multi-stage continual adaptation would move age-agnostic speaker verification closer to practical deployment.

9.4 Conclusion

This dissertation began with the observation that children’s speech changes substantially during development and that these changes create challenges for speaker verification. Through longitudinal acoustic and articulatory analysis, the dissertation quantified developmental variability in child speech. Through age-robust evaluation and adaptation methods, it showed that child speaker verification systems must explicitly address inter-year variability. Through adapter-based fine-tuning, it demonstrated that adult-trained models can be efficiently adapted to child speech under limited data. Through age-agnostic modeling and source-free adaptation, it addressed the broader challenge of improving child speaker verification while preserving adult-domain knowledge.

The main conclusion is that robust child speaker verification requires a developmental perspective. Systems must be designed not only to recognize who is speaking, but also to handle how a speaker’s voice changes with age. By connecting developmental speech analysis with domain adaptation, catastrophic forgetting, and source-free learning, this dissertation provides a foundation for speaker verification systems that are more inclusive, practical, and robust across age groups.

REFERENCES

- [1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN Embeddings for Speaker Recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [2] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in *Interspeech*, 2020, pp. 3830–3834.
- [3] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak *et al.*, “State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations,” *Computer Speech & Language*, vol. 60, p. 101026, 2020.
- [4] T. Zhou, Y. Zhao, and J. Wu, “Resnext and res2net structures for speaker verification,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 301–307.
- [5] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. Meng, “MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification,” in *Interspeech*, 2022, pp. 306–310.
- [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, 1999.
- [7] G. Yeung, S. M. Lulich, J. Guo, M. S. Sommers, and A. Alwan, “Subglottal resonances of American English speaking children,” *The Journal of the Acoustical Society of America*, vol. 144, no. 6, pp. 3437–3449, 2018.
- [8] S. Lee, A. Potamianos, and S. Narayanan, “Developmental acoustic study of American English diphthongs,” *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1880–1894, 2014.
- [9] H. K. Vorperian and R. D. Kent, “Vowel acoustic space development in children: A synthesis of acoustic and anatomic data,” *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 6, pp. 1510–1545, 2007.
- [10] G. E. Peterson and H. L. Barney, “Control methods used in a study of the vowels,” *The Journal of the acoustical society of America*, vol. 24, no. 2, pp. 175–184, 1952.
- [11] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, “Acoustic characteristics of American English vowels,” *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.
- [12] Y. Jung, “Acoustic articulatory evidence for quantal vowel categories: the features [low] and [back],” Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

- [13] K. N. Stevens, “On the quantal nature of speech,” *Journal of phonetics*, vol. 17, no. 1, pp. 3–45, 1989.
- [14] K. N. Stevens and S. J. Keyser, “Quantal theory, enhancement and overlap,” *Journal of phonetics*, vol. 38, no. 1, pp. 10–19, 2010.
- [15] S. M. Lulich, “Subglottal resonances and distinctive features,” *Journal of Phonetics*, vol. 38, no. 1, pp. 20–32, 2010.
- [16] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, “In-domain and out-of-domain data augmentation to improve children’s speaker verification system in limited data scenario,” in *ICASSP*. IEEE, 2020, pp. 7554–7558.
- [17] S. Aziz and S. Shahnawazuddin, “Enhancing Children’s Short Utterance-Based ASV Using Inverse Gamma-tone Filtered Cepstral coefficients,” *Circuits, Systems, and Signal Processing*, vol. 43, no. 5, pp. 3020–3041, 2024.
- [18] S. Aziz, Ankita, and S. Shahnawazuddin, “Short-Utterance-Based Children’s Speaker Verification in Low-Resource Conditions,” *Circuits, Systems, and Signal Processing*, vol. 43, no. 3, pp. 1715–1740, 2024.
- [19] V. P. Singh, M. Sahidullah, and T. Kinnunen, “ChildAugment: Data augmentation methods for zero-resource children’s speaker verification,” *The Journal of the Acoustical Society of America*, vol. 155, no. 3, pp. 2221–2232, 2024.
- [20] K. Radha, M. Bansal, and R. B. Pachori, “Automatic speaker and age identification of children from raw speech using sincNet over ERB scale,” *Speech Communication*, vol. 159, p. 103069, 2024.
- [21] S. Kataria, P. S. Nidadavolu, J. Villalba, and N. Dehak, “Analysis of Deep Feature Loss Based Enhancement for Speaker Verification,” in *The Speaker and Language Recognition Workshop (Odyssey)*, 2020, pp. 459–466.
- [22] M. Sang and J. H. Hansen, “Efficient Adapter Tuning of Pre-Trained Speech Models for Automatic Speaker Verification,” in *ICASSP*. IEEE, 2024, pp. 12 131–12 135.
- [23] W. Huang, B. Han, S. Wang, Z. Chen, and Y. Qian, “Robust Cross-Domain Speaker Verification with Multi-Level Domain Adapters,” in *ICASSP*. IEEE, 2024, pp. 11 781–11 785.
- [24] T. Wang, L. Li, and D. Wang, “SE/BN Adapter: Parametric Efficient Domain Adaptation for Speaker Recognition,” in *Interspeech*, 2024, pp. 2145–2149.
- [25] S. Han, Y. Ahn, K. Kang, and J. W. Shin, “Short-segment speaker verification using ecapa-tdnn with multi-resolution encoder,” in *ICASSP*, 2023, pp. 1–5.
- [26] Z.-C. Chen, Y.-S. Sung, and H.-y. Lee, “Chapter: Exploiting convolutional neural network adapters for self-supervised speech models,” in *ICASSP Workshops*, 2023, pp. 1–5.

- [27] J. Peng, T. Stafylakis, R. Gu, O. Plchot, L. Mošner, L. Burget, and J. Černocký, “Parameter-efficient transfer learning of pre-trained transformer models for speaker verification using adapters,” in *ICASSP*, 2023, pp. 1–5.
- [28] R. Li, W. Zhang, and D. Chen, “The coral++ algorithm for unsupervised domain adaptation of speaker recognition,” in *ICASSP*. IEEE, 2022, pp. 7172–7176.
- [29] W. Lin and M.-W. Mak, “Robust speaker verification using deep weight space ensemble,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 802–812, 2023.
- [30] J. Thienpondt, B. Desplanques, and K. Demuynck, “Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization,” in *Interspeech 2020*, 2020, pp. 756–760.
- [31] N. R. Koluguri, M. Kumar, S. H. Kim, C. Lord, and S. Narayanan, “Meta-learning for robust child-adult classification from speech,” in *ICASSP*. IEEE, 2020, pp. 8094–8098.
- [32] M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” *Psychology of learning and motivation*, 1989.
- [33] R. M. French, “Catastrophic forgetting in connectionist networks,” *Trends in Cognitive Sciences*, 1999.
- [34] S. Ruder, “An overview of multi-task learning in deep neural networks,” *arXiv preprint arXiv:1706.05098*, 2017.
- [35] S. Vander Eeckt and H. Van Hamme, “Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition,” in *ICASSP*, 2023.
- [36] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, “Progressive neural networks,” *arXiv preprint arXiv:1606.04671*, 2016.
- [37] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, “Lifelong Learning with Dynamically Expandable Networks,” in *ICLR*, 2018.
- [38] A. Mallya and S. Lazebnik, “Packnet: Adding multiple tasks to a single network by iterative pruning,” in *CVPR*. IEEE, 2018, pp. 7765–7773.
- [39] A. Robins, “Catastrophic forgetting, rehearsal and pseudorehearsal,” *Connection Science*, 1995.
- [40] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, “Overcoming catastrophic forgetting in neural networks,” *PNAS*, 2017.
- [41] F. Zenke, B. Poole, and S. Ganguli, “Continual learning through synaptic intelligence,” in *ICML*. PMLR, 2017, pp. 3987–3995.

- [42] A. Krogh and J. Hertz, “A simple weight decay can improve generalization,” *NeurIPS*, 1991.
- [43] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *NIPS*, 2014.
- [44] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, “Memory aware synapses: Learning what (not) to forget,” in *ECCV*. Springer, 2018, pp. 144–161.
- [45] H. Shin, J. K. Lee, J. Kim, and J. Kim, “Continual learning with deep generative replay,” *NeurIPS*, 2017.
- [46] Z. Li and D. Hoiem, “Learning without forgetting,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.
- [47] S. Charles and S. M. Lulich, “An ultrasound study of American English laterals produced by children,” in *Proceedings of Meetings on Acoustics*, vol. 36, no. 1. AIP Publishing, 2019.
- [48] S. M. Lulich and S. D. Charles, “Development of formant frequency distributions in American English-speaking elementary school-aged children: A longitudinal study,” in *Proceedings of Meetings on Acoustics 179ASA*, vol. 42, no. 1. Acoustical Society of America, 2020, p. 060016.
- [49] K. Shobaki, J.-P. Hosom, and R. Cole, “The OGI kids’ speech corpus and recognizers,” in *Proc. of ICSLP*. Citeseer, 2000, pp. 564–567.
- [50] W. Ward, R. Cole, and S. Pradhan, “My science tutor and the myst corpus,” *Boulder Learning Inc*, 2019.
- [51] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A Large-Scale Speaker Identification Dataset,” in *Interspeech*, 2017, pp. 2616–2620.
- [52] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep Speaker Recognition,” in *Interspeech*, 2018, pp. 1086–1090.
- [53] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, “Cn-celeb: a challenging chinese speaker recognition dataset,” in *ICASSP*. IEEE, 2020, pp. 7604–7608.
- [54] L. Ménard, J. Aubin, M. Thibeault, and G. Richard, “Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model,” *Folia Phoniatrica et Logopaedica*, pp. 64–72, 2012.
- [55] N. Zharkova, F. E. Gibbon, and W. J. Hardcastle, “Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation,” *Clinical Linguistics & Phonetics*, pp. 249–265, 2015.

- [56] N. Zharkova, F. E. Gibbon, and A. Lee, “Using ultrasound tongue imaging to identify covert contrasts in children’s speech,” *Clinical Linguistics & Phonetics*, pp. 21–34, 2017.
- [57] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Interspeech*, 2015, pp. 3214–3218.
- [58] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2net: A new multi-scale backbone architecture,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 652–662, 2019.
- [59] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *CVPR*, 2018, pp. 7132–7141.
- [60] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive Statistics Pooling for Deep Speaker Embedding,” in *Interspeech*, 2018, pp. 2252–2256.
- [61] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.
- [62] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” in *Interspeech*, 2015, pp. 3586–3589.
- [63] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *ICASSP. IEEE*, 2017, pp. 5220–5224.
- [64] 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,” in *Interspeech*, 2019, pp. 2613–2617.
- [65] S. M. Lulich, J. R. Morton, H. Arsikere, M. S. Sommers, G. K. Leung, and A. Alwan, “Subglottal resonances of adult male and female native speakers of American English,” *The Journal of the Acoustical Society of America*, pp. 2592–2602, 2012.
- [66] T. G. Csapó, Z. Bárkányi, T. E. Grácz, T. Böhm, and S. M. Lulich, “Relation of formants and subglottal resonances in Hungarian vowels,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.
- [67] X. Chi and M. Sonderegger, “Subglottal coupling and its influence on vowel formants,” *The Journal of the Acoustical Society of America*, pp. 1735–1745, 2007.
- [68] G. Dogil, S. M. Lulich, A. Madsack, and W. Wokurek, “Crossing the quantal boundaries of features: subglottal resonances and Swabian diphthongs,” *Tones and Features: Phonetic and Phonological Perspectives. De Gruyter Mouton*, pp. 137–148, 2011.
- [69] L. Ménard, “Acoustic variability and adaptive articulatory strategies during vocal tract growth revealed by the rounding contrast in french,” in *15th International Congress of Phonetic Sciences*, 2003, pp. 3169–3172.

- [70] K. Stevens, “Acoustic phonetics,” 1999.
- [71] U. G. Goldstein, “An articulatory model for the vocal tracts of growing children,” Ph.D. dissertation, Massachusetts Institute of Technology, 1980.
- [72] A. O. Santos, J. Godoy, K. Silverio, and A. Brasolotto, “Vocal changes of men and women from different age decades: an analysis from 30 years of age,” *Journal of Voice*, pp. 840–850, 2023.
- [73] X. Qin, N. Li, W. Chao, D. Su, and M. Li, “Cross-Age Speaker Verification: Learning Age-Invariant Speaker Embeddings,” in *Interspeech*, 2022, pp. 1436–1440.
- [74] G. Yeung and A. Alwan, “On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children,” in *Interspeech*, 2018, pp. 1661–1665.
- [75] V. M. Shetty, S. M. Lulich, and A. Alwan, “Developmental Articulatory and Acoustic Features for Six to Ten Year Old Children,” in *Interspeech*, 2023, pp. 4598–4602.
- [76] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in neural information processing systems*, pp. 17 022–17 033, 2020.
- [77] <https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/SpeechSynthesis/HiFiGAN>.
- [78] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.
- [79] 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 *Interspeech*, 2022, pp. 4900–4904.
- [80] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *International conference on machine learning*. PMLR, 2017, pp. 933–941.
- [81] W. F. Katz and P. F. Assmann, “Identification of children’s and adults’ vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing,” *Journal of Phonetics*, pp. 23–51, 2001.
- [82] S. Ghai and R. Sinha, “Exploring the effect of differences in the acoustic correlates of adults’ and children’s speech in the context of automatic speech recognition,” *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1–15, 2010.
- [83] M. J. Russell and S. D’Arcy, “Challenges for computer recognition of children’s speech.” *SLaTE*, p. 111, 2007.

- [84] J. Zhang, J. Ma, X. Guo, L. Li, and L. He, “A speaker recognition method based on stable learning,” in *ICASSP*. IEEE, 2024, pp. 10 221–10 225.
- [85] L. Sari, S. Thomas, M. Hasegawa-Johnson, and M. Picheny, “Pre-training of speaker embeddings for low-latency speaker change detection in broadcast news,” in *ICASSP*. IEEE, 2019, pp. 6286–6290.
- [86] R. A. Fisher, “On the mathematical foundations of theoretical statistics,” in *Breakthroughs in statistics: Foundations and basic theory*. Springer, 1992, pp. 11–44.