Robust Child Speech Classification Leveraging Augmentation and Speaker Embeddings

Pranav Bijith
Santa Susana High School
Simi Valley, United States
pranavbijith1@gmail.com

Sebastian Farrington
Arcadia High School
Arcadia, United States
sebastian.farrington@gmail.com

Alexander Gardiner
Dos Pueblos High School
Goleta, United States
misc@alexandergardiner.com

Andrew Liu
Carlmont High School
Belmont, United States
ajliu2016@gmail.com

Leonard Collomb
Lick-Wilmerding High School
San Francisco, United States
le.collomb@gmail.com

***Abstract*—Classifying child speech is challenging due to the limited availability of large, publicly accessible datasets recorded in consistent environments. This study aimed to develop a classifier to differentiate child and adult speech using datasets such as LibriSpeech, VoxCeleb, Common Voice, OGI Kids, CMU Kids, and MyST. Initial attempts using diarization and embeddings from ECAPA-TDNN, Whisper, and HuBERT led to misclassification due to dataset-specific clustering caused by recording artifacts. To address this, data augmentation with impulse responses was applied, improving generalization. A fine-tuned gradient boosting classifier trained on ECAPA-TDNN embeddings achieved 93.5% accuracy. Future efforts will explore enhanced augmentation and model tuning for further improvements.**

Keywords—Child Speech Classification, Embeddings, Speech Processing, HuBERT, Whisper, SpeechBrain, Gradient Boosting, Data Augmentation

I. Introduction

Classifying child speech is a persistent challenge in computational linguistics and speech processing, largely due to the legal, ethical, and technical difficulties of acquiring large, publicly available datasets, as well as the intrinsic variability of child speech. This limitation hinders the development of automatic speech recognition (ASR) systems and voice activity detection (VAD) models optimized for children, as most are predominantly trained on adult speech, leading to reduced performance on child speech. Smaller datasets lack the representativeness required to capture the diversity of child speech patterns, further compounding the problem [1]. While humans excel at distinguishing child and adult speech based on age-related vocal characteristics, research shows that identifying individual child speakers remains challenging even for trained adults, highlighting the inherent difficulty of the task [2].

Accurate classification of child speech is critical for applications in speech therapy, education, and voice-based interfaces. For instance, automated scoring systems in speech therapy can complement clinical services, enabling higher-intensity practice and improved outcomes [3]. Similarly, adaptive educational tools can personalize learning experiences for children. The absence of robust child speech detection solutions thus presents a significant barrier to advancing these technologies.

This study proposes an approach to enhance child-adult speech classification by leveraging publicly available datasets and mitigating dataset-specific biases through data augmentation and embedding-based modeling. The methodology incorporates speaker diarization, impulse response augmentation, and embedding extraction using models such as ECAPA-TDNN, Whisper, and HuBERT. Diverse machine learning techniques, including neural networks, logistic regression, and gradient boosting, are explored, with performance evaluated using metrics such as accuracy, recall, precision, and F1 scores. Visualization techniques like t-SNE further elucidate the separability of child and adult speech embeddings. The proposed framework achieves up to 93.5% classification accuracy on augmented datasets, offering a robust solution for child speech classification despite the variability inherent in child speech.

II. Previous Methods

Several methods have been proposed to address the challenges of child-adult speech classification. Koluguri et al. utilized meta-learning with prototypical neural networks, treating each child-adult pair as a separate task during training. This approach improved generalizability over supervised learning methods, yielding a 14.53% improvement in F1 scores and a 9.66% gain in cluster purity. Their model achieved a peak F1 score of 0.8555 using protonets with x-vectors [4]. Lahiri et al. tackled variability in developmental aspects and background conditions using domain adversarial training, extracting embeddings with generative adversarial networks (GANs), and gradient reversal. Their score fusion approach yielded F1 scores ranging from 76.85% to 85.21% across different age groups [5]. Both studies demonstrated promising advancements but faced limitations in generalizing to unseen datasets.

Other works explored diverse approaches to improve speech classification. Serizel and Giuliani combined deep neural networks (DNNs) and hidden Markov models (HMMs) to reduce phone error rates for Italian datasets, achieving up to 92.9% accuracy [6]. Xu et al. proposed a multimodal method incorporating audio and visual features for speaker classification, leveraging models like Whisper and ResNet for utterances lasting 0.3–3.0 seconds. This approach achieved macro F1 scores of up to 93.6% [7]. Lastly, Abed and Sztaho investigated deep embedding methods such as ECAPA-TDNN for Icelandic speech, demonstrating high accuracy for older speakers but reduced performance for children [8]. In contrast, our work focuses exclusively on English speech, employing robust augmentation techniques and advanced embedding models to achieve 93.5% classification accuracy, addressing the limitations of prior studies.

III. Methodology

A. Datasets

This study utilized six datasets: three for adult speech (LibriSpeech, VoxCeleb, and Common Voice) and three for child speech (OGI Kids, CMU Kids, and MyST). LibriSpeech, an ASR corpus derived from public-domain audiobooks, provided clean, high-quality adult speech data, with the "train.100" split used for training [9]. VoxCeleb, featuring speech from over 7,000 speakers of diverse backgrounds, served as an adult testing dataset [10]. Common Voice, a crowdsourced corpus of validated voices, was filtered to include adult speakers aged 20–50 and used for training [11].

For child speech, OGI Kids included kindergarten-aged scripted and spontaneous utterances, with 1,101 samples selected after preprocessing [12]. CMU Kids, consisting of read-aloud sentences by children aged 6–11, contributed 5,180 utterances to the training set [13]. Lastly, the MyST dataset provided approximately 470 hours of English speech from students in grades 3–5, used exclusively for testing [14]. Together, these datasets ensured diverse and representative speech samples for robust child-adult classification.

Diarization was used to segment audio recordings into speaker-specific utterances, increasing the number of usable samples for training and improving model accuracy. The pre-trained model pyannote/speaker-diarization@2.1, known for its three-stage pipeline and adaptability, was employed for this purpose [16]. Fine-tuning focused on the collar parameter, set to 0.8 to account for the frequent pauses in child speech and ensure accurate segmentation. All audio files were diarized, and the resulting timestamps were stored for further processing, significantly enhancing the dataset's utility for classification.

Table 1: Details of available utterances per dataset post-diarization

Dataset	Utterances after Diarizations
OGI Kids	14117
CMU Kids	3857
MyST	4063
Librispeech	4098
VoxCeleb	3031
Common Voice	2770

B. Augmentation

Initial model training yielded low accuracies, with predictions skewed entirely toward either child or adult speech. Analysis with t-SNE revealed that embeddings were heavily influenced by spatial audio characteristics, leading to distinct but non-overlapping clusters for each dataset. This lack of overlap between child and adult speech clusters indicated that spatial audio differences, rather than inherent speech features, were driving classification. To address this issue, impulse response augmentation was applied to normalize spatial audio across datasets, enabling the model to focus on distinctive features relevant to child-adult classification.

Impulse responses, which mimic reverberation, echo, and filtering effects, were sourced from the MIT Acoustical Reverberation Scene Statistics Survey [15]. Four high-magnitude impulse responses—representing environments such as a bar, stairwell, and supermarket—were selected to ensure strong augmentation. Using Scipy's signal processing library, 3,000 audio samples from each dataset (except OGI Kids) were augmented with each impulse response, creating four new datasets per original dataset. For OGI Kids, all 1,101 samples were similarly augmented. This augmentation strategy harmonized spatial audio settings across datasets, enhancing model robustness and improving classification performance.

C. Embeddings

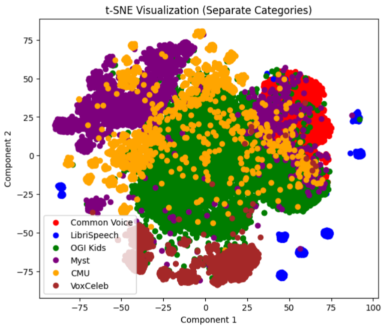
This research utilized three types of audio embeddings: ECAPA-TDNN, Whisper, and HuBERT, each applied to diarized audio segments across all 24 datasets. HuBERT (HuBERT-base-ls960), a self-supervised model, leverages transformer-based architecture to learn phonetic and prosodic speech representations. Using the Wave2Vec hubert-large-ls960-ft processor, it extracts 768 features per audio file, making it suitable for capturing deep acoustic patterns [17]. Whisper, a state-of-the-art embedding extractor, employs a transformer encoder-decoder trained on 680,000 hours of multilingual speech data. It directly processes raw audio files and extracts 768 features, enabling its use in various speech tasks, including ASR [18].

ECAPA-TDNN is a robust embedding extractor employing unique mechanisms like Squeeze-and-Excitation Blocks, Res2Net modules, and attentive statistics pooling to enhance feature granularity and noise resilience. Unlike Whisper and HuBERT, it generates 192 features per audio file, optimized for speaker-relevant signal encoding [19]. All embeddings were computed for diarized segments and stored as NumPy files for every dataset, providing comprehensive feature representations for subsequent model training and analysis. These embeddings are compared in terms of performance and suitability for different model architectures in the conclusion.

D. Visualization

Embeddings were visualized using t-SNE, a technique for reducing high-dimensional data into two dimensions, allowing clusters to be easily analyzed. This method highlighted distinctions and overlaps between child and adult speech data, providing insights into the separability of the two categories and the model's potential applicability to real-world scenarios. Three types of t-SNE visualizations were created: individual datasets were color-coded in the first, adult and child datasets were grouped and colored in the second, and the third used transparency to analyze overlap between adult and child data.

Fig 1: t-SNE plot for Whisper embeddings from different datasets



E. Models/Training

This study utilized four types of models—neural networks, logistic regression, transformers and the gradient boosting algorithm XGBoost to classify child and adult speech. The training datasets included OGI Kids and CMU Kids for child speech and Common Voice and LibriSpeech for adult speech, while MyST and VoxCeleb were used for testing child and adult speech, respectively. To enhance model robustness, 16 augmented datasets were created from the original training datasets, each representing one of four impulse response augmentations. Utterances were randomized and labeled based on their dataset of origin, and class balancing was achieved by trimming excess child utterances.

Transformers, neural networks, and logistic regression were tested for their ability to distinguish child from adult speech. Transformers, built using the AutoModel architecture, directly processed embeddings with a simplified architecture, including one fully connected layer and a sequence layer to manage temporal information. Neural networks with three layers were effective due to the clear separability of child and adult speech clusters observed in t-SNE visualizations. Hyperparameter tuning, including adjustments to learning rate and batch size, further optimized neural network performance. Logistic regression, leveraging its simplicity, captured straightforward distinctions in the embeddings and proved reliable for this task.

Gradient boosting, implemented via XGBoost, delivered the best overall performance, achieving a classification accuracy of 93.5%. This method combined successive weak models with continuous weight corrections to improve predictions. Initial parameter grid searches led to overfitting; however, refined parameters yielded higher test accuracies while maintaining generalizability. XGBoost's robustness against overfitting and its capacity to handle diverse augmented datasets made it the most effective model for child-adult speech classification in this study.

IV. Results

A. Model Performance

The evaluation metrics for this study included accuracy, precision, recall, and F1 score. High recall indicated effective child speech prediction, while high precision reflected accurate adult speech classification. The F1 score, a harmonic mean of precision and recall, served as a comprehensive measure of overall model performance, mitigating the impact of class imbalances.

Various models were tested, including neural networks, logistic regression, transformers, and gradient boosting, using embeddings from ECAPA-TDNN, HuBERT, and Whisper.ECAPA-TDNN consistently outperformed other embeddings due to its focus on speaker-specific features, achieving the highest accuracy and F1 scores across all models. The best neural network model with ECAPA-TDNN achieved an F1 score of 0.9172, outperforming similar models with Whisper and HuBERT embeddings. However, neural networks demonstrated lower precision for adult speech predictions, indicating room for improvement in balancing class-specific performance.

Logistic regression performed strongly with ECAPA-TDNN embeddings, achieving an accuracy of 92.17% and an F1 score of 0.9253. Its simplicity allowed it to focus effectively on key features, avoiding overfitting to non-voice-specific features, which limited Whisper and HuBERT's performance. Despite its limitations in flexibility compared to neural networks, logistic regression proved highly effective with ECAPA-TDNN embeddings.

Transformers exhibited mixed results. Whisper and HuBERT embeddings, when paired with transformers, often overfitted at higher learning rates. Adjustments to hyperparameters, such as reducing learning rates and limiting epochs to five, improved their performance. Whisper embeddings paired with transformers achieved an F1 score of 0.9036, but ECAPA-TDNN embeddings again proved the most robust, achieving an F1 score of 0.9218, demonstrating its adaptability to transformer architectures.

Gradient boosting with XGBoost emerged as the most effective model overall. Using ECAPA-TDNN embeddings and optimized parameters, it achieved the highest accuracy of 93.52% and an F1 score of 93.81%. Its robustness in preventing overfitting enabled it to generalize well across testing datasets, outperforming all other models. While Whisper and HuBERT embeddings showed incremental improvements in certain cases, they fell short of ECAPA-TDNN embedding's superior performance, highlighting the importance of embedding quality for gradient boosting's success.

Table 2: Performance of HuBERT embeddings across different models

Model	Accuracy	F-1 Score
Neural Networks	0.9035	0.9102
Logistic Regression	0.8688	0.8767
Transformer	0.8556	0.8709
XGBoost	0.8805	0.8910

Table 3: Performance of Whisper embeddings across different models

Model	Accuracy	F-1 Score
Neural Networks	0.8924	0.9018
Logistic Regression	0.8429	0.8629
Transformer	0.8978	0.9036
XGBoost	0.9035	0.9101

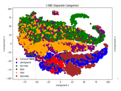
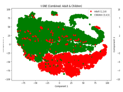
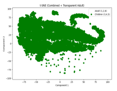
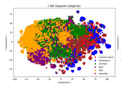
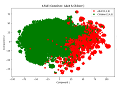
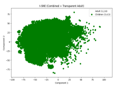
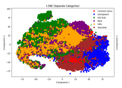
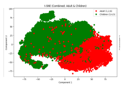
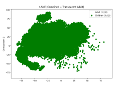
Table 4: Performance of ECAPA-TDNN embeddings across different models

Model	Accuracy	F-1 Score
Neural Networks	0.9132	0.9172
Logistic Regression	0.9217	0.9253
Transformer	0.9192	0.9218
XGBoost	0.9352	0.9381

B. T-SNE Visual Data

The T-SNE visualizations display embeddings based on embedding type and impulse response used for augmentation. Plots include individual datasets, combined child and adult datasets, and child data overlaid with transparent adult data to assess overlap. Dataset colors are OGI Kids (green), MyST (purple), CMU Kids (yellow), VoxCeleb (maroon), LibriSpeech (blue), and Common Voice (red). In combined plots, child speech is green, and adult speech is red or transparent.

Table 5: t-SNE plots with augmentation for different embeddings

Embeddings	T-SNEs (All Datasets Separate)	T-SNEs (Child and Adult)	T-SNEs (Child + Adult Transparent)
Whisper			
ECAPA-TDNN			
Hubert			

V. Conclusions

This study demonstrated the effectiveness of various machine learning models and embeddings for child-adult speech classification. Among the approaches tested, gradient boosting with ECAPA-TDNN embeddings emerged as the most robust, achieving the highest accuracy and F1 score, while logistic regression and neural networks provided valuable insights into the impact of model simplicity and hyperparameter tuning. The addition of impulse response augmentation proved critical in normalizing spatial audio features and improving classification performance across all models. Visual analyses using T-SNE highlighted clear separations between child and adult speech clusters, further validating the methodology. However, the study also revealed the limitations of embeddings like Whisper and HuBERT, which struggled with overfitting due to their focus on non-voice-specific features, emphasizing the importance of embedding quality in determining model success.

To further advance this work, future research should focus on diversifying the training data through more extensive impulse response augmentation, utilizing all 271 responses from the MIT survey. Incorporating additional datasets and leveraging larger training samples could enhance model generalizability and real-world applicability. Transitioning to unaugmented test datasets would better validate the robustness of these models in practical settings. While this study highlights the potential of ECAPA-TDNN embeddings and gradient boosting for child-adult speech classification, addressing computational constraints and expanding data diversity are key to unlocking even greater performance. These advancements could make significant contributions to applications such as speech therapy, assistive learning, and voice-driven user interfaces, where reliable child speech detection is essential.

References

[1] Yu, Fan, et al. “The SLT 2021 Children Speech Recognition Challenge: Open.Datasets, Rules and Baselines.” ArXiv.org, 2021, arxiv.org/abs/2011.06724?utm_source=chatgpt.com. Accessed 16 Dec. 2024.

[2] Cooper, Angela, et al. “Identifying Children’s Voices.” The Journal of the Acoustical Society of America, vol. 148, no. 1, 1 July 2020, pp. 324–333, https://doi.org/10.1121/10.0001576. Accessed 16 Dec. 2024.

[3] Mcallister, Tara, et al. PERCEPT: A Database of Clinical Child Speech for Automatic Speech Recognition and Classification.

[4] Rao, Koluguri Nithin, et al. “Meta-Learning for Robust Child-Adult Classification from Speech.” ArXiv.org, 2019, arxiv.org/abs/1910.11400. Accessed 16 Dec. 2024.

[5] Lahiri, Rimita, et al. “Learning Domain Invariant Representations for Child-Adult Classification from Speech.” ArXiv.org, 24 Oct. 2019, arxiv.org/abs/1910.11472.

[6] Serizel, Romain, and Diego Giuliani. Deep Neural Network Adaptation for Children’s and Adults’ Speech Recognition Deep Neural Network Adaptation for Children’s and Adults’ Speech Recognition. 16 Aug. 2016.

[7] Xu, Anfeng, et al. “Audio-Visual Child-Adult Speaker Classification in Dyadic Interactions.” ArXiv.org, 2023, arxiv.org/abs/2310.01867. Accessed 16 Dec. 2024.

[8] Abed, Mohammed Hamzah, and Dávid Sztahó. “Deep Speaker Embeddings for Speaker Verification of Children.” Lecture Notes in Computer Science, 1 Jan. 2024, pp. 58–69, https://doi.org/10.1007/978-3-031-70566-3_6. Accessed 16 Dec. 2024.

[9] Panayotov, Vassil, et al. “Librispeech: An ASR Corpus Based on Public Domain Audio Books.” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2015, https://doi.org/10.1109/icassp.2015.7178964.

[10] Nagrani, Arsha, et al. “VoxCeleb: Large-Scale Speaker Verification in the Wild.” Computer Speech & Language, Oct. 2019, p. 101027, www.robots.ox.ac.uk/~vgg/publications/2017/Nagrani17/nagrani17.pdf, https://doi.org/10.1016/j.csl.2019.101027.

[11] “Common Voice by Mozilla.” Commonvoice.mozilla.org, commonvoice.mozilla.org/en.

[12] Shobaki, Khaldoun, et al. “The OGI Kids2 Speech Corpus and Recognizers.” 6th International Conference on Spoken Language Processing (ICSLP 2000), 16 Oct. 2000, https://doi.org/10.21437/icslp.2000-800. Accessed 8 Oct. 2022.

[13] Eskenazi, Maxine, et al. “The CMU Kids Corpus - Linguistic Data Consortium.” Upenn.edu, 2019, catalog.ldc.upenn.edu/LDC97S63.

[14] Pradhan, Sameer, et al. “MyST Children’s Conversational Speech - Linguistic Data Consortium.” Upenn.edu, 2021, catalog.ldc.upenn.edu/LDC2021S05.

[15] Traer, James , and Josh McDermott. “IR Survey (Traer and McDermott).” Mcdermottlab.mit.edu, 2016, mcdermottlab.mit.edu/Reverb/IR_Survey.html.

[16] Bredin, H. (2023) pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. Proc. INTERSPEECH 2023, 1983-1987, doi: 10.21437/Interspeech.2023-105

[17] Hsu, Wei-Ning, et al. “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.” ArXiv:2106.07447 [Cs, Eess], 14 June 2021, arxiv.org/abs/2106.07447.

[18] Radford, Alec, et al. Robust Speech Recognition via Large-Scale Weak Supervision. 6 Dec. 2022.

[19] Desplanques, Brecht, et al. “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification.” Interspeech 2020, 25 Oct. 2020, pp. 3830–3834, arxiv.org/abs/2005.07143, https://doi.org/10.21437/Interspeech.2020-2650.