



Augmentation techniques for adult-speech to generate child-like speech data samples at scale

Title	Augmentation techniques for adult-speech to generate child-like speech data samples at scale
Author(s)	Yahayah Yiwere, Mariam;Barcovschi, Andrei;Jain, Rishabh;Cucu, Horia;Corcoran, Peter
Publication Date	2023-09-20
Publisher	IEEE
Repository DOI	https://doi.org/10.1109/ACCESS.2023.3317360

RESEARCH ARTICLE

Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale

MARIAM YAHAYAH YIWERE¹, ANDREI BARCOVSKI¹,
RISHABH JAIN¹, (Graduate Student Member, IEEE), HORIA CUCU², (Member, IEEE),
AND PETER CORCORAN¹, (Fellow, IEEE)

¹School of Electrical and Electronics Engineering, University of Galway, Galway, H91 TK33 Ireland

²Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, 060042 Bucharest, Romania

Corresponding author: Mariam Yahayah Yiwere (mariam.yiwere@universityofgalway.ie)

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

ABSTRACT Technologies such as Text-To-Speech (TTS) synthesis and Automatic Speech Recognition (ASR) have become important in providing speech-based Artificial Intelligence (AI) solutions in today's AI-centric technology sector. Most current research work and solutions focus largely on adult speech compared to child speech. The main reason for this disparity can be linked to the limited availability of children's speech datasets that can be used in training modern speech AI systems. In this paper, we propose and validate a speech augmentation pipeline to transform existing adult speech datasets into synthetic child-like speech. We use a publicly available phase vocoder-based toolbox for manipulating sound files to tune the pitch and duration of the adult speech utterances making them sound child-like. Both objective and subjective evaluations are performed on the resulting synthetic child utterances. For the objective evaluation, the similarities of the selected top adults' speaker embeddings are compared before and after the augmentation to a mean child speaker embedding. The average adult voice is shown to have a cosine similarity of approximately 0.87 (87%) relative to the mean child voice after augmentation, compared to a similarity of approximately 0.74 (74%) before augmentations. Mean Opinion Score (MOS) tests were also conducted for the subjective evaluation, with average MOS scores of 3.7 for how convincing the samples are as child-speech and 4.6 for how intelligible the speech is. Finally, ASR models fine-tuned with the augmented speech are tested against a baseline set of ASR experiments showing some modest improvements over the baseline model finetuned with only adult speech.

INDEX TERMS Adult speech datasets, child speech datasets, synthetic child speech, speech data augmentation, CLEESE, speaker embeddings, pitch tuning, fundamental frequency.

I. INTRODUCTION

In recent years, rapid advances in Machine Learning (ML) and Deep Neural Network (DNN) techniques, together with tremendous increases in computational power, have led to a significant boost in the development of speech related

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik¹.

technologies such as ASR [1], [2], [3], [4], [5], [6], [7], [8], TTS [9], [10], [11], [12], [13], [14] and Speaker Recognition [15], [16], [17], [18], [19] for multiple application domains. However, most of the solutions to date focus largely on adult speech, leading to poor performance when dealing with children's speech. The relatively smaller amount of work done specifically with child speech [20], [21], [22], [23], [24], [25], [26] has encountered significant challenges,

and as a result, children cannot fully benefit from some modern speech technologies. One of the main challenges is the limited availability of children's speech datasets [27], [28], [29], [30] necessary and suitable for training speech AI models. This is in contrast to adult speech where there is an abundance of large, publicly available and well-annotated datasets [31], [32], [33], [34], [35], [36]. These are harvested from the vast amounts of high-quality public data available online, including YouTube videos and professionally recorded audiobooks. Unfortunately, these adult datasets are not suitable for developing child-friendly speech AI solutions due to the innate differences between child and adult speech.

Child speech differs in multiple ways from adult speech owing essentially to the anatomical and morphological differences in their vocal-tract structure. Children have shorter vocal cords, giving their voices higher fundamental and formant frequencies compared to adults. In addition, children may have less control over articulation and non-linguistic aspects of speech such as prosody and therefore child speech exhibits higher spectral and temporal variation than adult speech [39].

On average, children also have slower speaking rates due to having longer phoneme durations [40]. They also exhibit higher pitch values: typically above 250Hz compared to average pitch values of 130Hz for adult males and 220Hz for adult females [41], [42]. For these reasons, it is important to gather and prepare good quality children's speech data to successfully train child-friendly speech-related AI models. However, there are additional challenges in the process of collecting child speech data [43], explaining the limited number of child-speech datasets available for research purposes.

A. EXISTING CHILD SPEECH DATASETS – DEFICIENCIES

There are some English child-speech datasets publicly available to researchers. Some of these [28], [29] were built using the approach of recruiting child speakers for recording sessions in professional recording studios, while others, for example, the MyST dataset [30] were built using a tablet or smartphone based app to record children's conversational speech remotely. For the latter, audio quality is highly dependent on the consumer device that the app runs on. All of these datasets feature several drawbacks, which affect data quality and introduces challenges to the use of said data in training speech-related AI models such as ASR and TTS. Invariably, major cleaning, filtering, annotation and other pre-processing of the data becomes necessary. A summary of the statistics and pros and cons of these child-speech datasets are presented in Table 1, along with some adult speech datasets for comparison.

A common problem with many of the child speech datasets is that they are relatively small/short in duration, as can be seen in Table 1, and are simply not enough in terms of duration (hours) to train a speech model on their own.

Another problem is the poor quality of recorded speech samples. Some datasets are generally of poor quality due to the recording devices and/or environments used to while capturing the data; for examples audio samples may have too much background noise, noise from recording gear, or very low gain. Lastly, some datasets have several bad speech samples.

For instance, in Table 1, MyST is the largest child dataset and has a lot of data (approx. 393 hours) from multiple speakers; but many of the utterances are too short or too long, non-meaningful or indiscernible and noisy. In addition, much of the dataset is not annotated, or annotations are of poor quality and cannot be used for training speech models [24].

B. CHALLENGES IN BUILDING CHILD-SPEECH DATASETS

Building a clean speech dataset even for adults is not an easy task. It requires a specially prepared environment (recording studio), the right recording and storage devices, as well as recruitment of speakers. Child speech data can also be collected using this traditional method of recruiting speaking actors for recording sessions in media studios; however, in the case of children, additional difficulties are introduced.

- Recruitment and data protection: The processes of recruiting child speakers (actors) and complying with data protection laws can be both expensive and time-consuming and must involve the parents or legal guardians of the children, as children cannot give their own legal consent.
- Low concentration and short attention spans: children have relatively lower levels of focus and shorter spans of attention, which could cut recording sessions short.
- Poor acoustic and linguistic capabilities of the youngest group of children.
- Poor quality of recording devices and environment.

Another approach that can be used to gather children's speech involves collecting audio recordings from the Internet, for example from YouTube or through a dedicated recording application. With this approach, a different set of challenges are faced:

- Limited number of videos with children as main actors.
- Short video/utterance durations.
- Background noise and music
- Lack of transcriptions and annotations.

C. RATIONALE FOR THIS RESEARCH

Taking all the above challenges into consideration, there is a need for alternative ways to build larger child speech datasets to facilitate the development of child-friendly speech technologies. To this end, the goal of this study is to explore the potential of augmenting adult speech to provide additional child-like speech samples to complement existing child-speech datasets. The resulting synthetic child voices can be used to generate more synthetic child speech with the appropriate (child-like) linguistic content using a fine-tuned TTS model.

TABLE 1. Summary of child speech research datasets with statistics & pros and cons.

Dataset	Main Statistics			Pros	Cons
	Type	Duration (hrs)	No. Speakers		
My Science Tutor (MyST) [30]	Child (grades 3-5)	393	1371	Large amount of data	Noisy (both audio & transcripts) Not fully transcribed
PF-STAR [27] (German, Italian, English Swedish)	Child	Approx. 65	611	Clean Fully transcribed	Small in size
CMU Kids [29]	Child (Ages 6-11)	Approx. 8.9	76	Clean Fully transcribed	Small in size
CSLU [28]	Child (grades 0-10)	100	1178	Transcribed (Scripted and spontaneous)	Noisy Extremely short utterances (single words)
LJ Speech [31]	Adult (female)	25	1	Clean, fully transcribed	Small in size
Librispeech [32]	Adult (male & female)	982	2,484	Large amount of data, fully transcribed	-
LibriTTS [37]	Adult (male & female)	586	2,456	Large amount of data, clean, fully transcribed, TTS-ready	-
VCTK [38]	Adult (male & female)	44	109	Clean, fully transcribed	Small in size

D. RELATED WORKS

To improve the performance of ASR models for children's speech, some researchers have adopted similar data augmentation techniques. For example, Shahnawazuddin et al. [44] proposed a prosody modification (i.e., pitch and speaking rate scaling) using a Zero-Frequency Filtering based Glottal Closure Instants (ZFF-GCI) anchoring approach. The authors used these modifications to introduce more variability in order to achieve speaker independent ASR and reported improvements in accuracy over their baseline for both adult and child test sets (ASR).

Bhardwaj et al. [45] also used pitch and speaking rate modification to improve performance of Punjabi ASR system on children's speech. It uses the ZFF-GCI method for Linear Prediction based Pitch Synchronous Overlap and Add (LP-PSOLA) together with speaker adaptive training and achieves an improvement in recognition rate for Punjabi child speech. Chen et al. [46] applied multiple modifications including pitch, tempo, speed, and volume perturbations to both adult and child training datasets to diversify and increase the amount of available training data to improve child ASR.

The idea of generating synthetic child-like speech from adult speech was explored by Singh et al. [47]. In their work, they applied spectral modifications, namely Linear Predicting Coding (LPC)-based segmental warping perturbations

(LPC-SWP) and formant energy perturbations (FEP), to adult data to generate child-like speech for data augmentation, and demonstrated an improvement in WER on both children and adult test sets when these modifications were combined with vocal tract length perturbation (VTLP).

Most of these works used different algorithmic approaches to apply prosody-based modifications (pitch and speaking rate scaling) to the speech, and the modifications were applied in a somewhat randomized manner. That is, both increasing and decreasing adjustments were applied to the audio features (e.g., pitch and speaking rate). In addition, the quality of the modified speech generated was not assessed in detail.

In this work, the goal is to generate/create synthetic child-like speech data, and we consider augmenting the pitch and speaking rate of adult speech to achieve this using a publicly available phase-vocoder based sound manipulation tool. To determine the timestamps of words and spaces where the speaking rate should be reduced, a forced alignment system based on an ASR model is used. In addition, we employ a speaker encoder model to visualize and compare the adults' and children's speaker embeddings in a common latent space before and after modifications. The contributions of this paper are as follows: a) exploring an alternative algorithm approach for the modification of adult speech (to make them more child-like through pitch and speaking rate adjustments),

b) conducting Mean Opinion Score (MOS) studies to provide a qualitative evaluation of the augmented/modified speech, c) scaling the augmentation to generate large amounts of synthetic child-like speech, d) conducting a proof-of-concept ASR experiment (example application) to provide a quantitative evaluation of the augmented adult speech.

The rest of this paper is organized as follows: Section II presents foundation technologies used in this research. Section III describes the methodology and Section IV presents the experiments conducted. Results and discussions are presented in Section V. Section VI shows an example application, and finally, Section VII presents our conclusions and future work.

II. FOUNDATION TOOLS AND TECHNOLOGIES

To develop our augmentation pipeline, we need to use a number of specialized tools to modify the pitch and control the duration of speech samples. In this section we introduce these tools, outline their features and discuss their role in the pipeline.

Different tools were considered for the tasks defined. The Combinatorial Expressive Speech Engine (CLEESE) [48] was selected to implement these augmentations because it offers a combination of ease-of-use and flexibility by allowing transformations to be applied to specific segments of the input speech sample where desired.

A. THE COMBINATORIAL EXPRESSIVE SPEECH ENGINE (CLEESE)

CLEESE is a python toolkit that can be used to perform deterministic or random transformations on input sound. Several features of the input sound can be modified, including the pitch, duration, and gain (amplitude). Originally designed to generate many random variations of a single input sound, CLEESE can also be used to perform individual and user-determined transformations, and the transformations can be either static or time-varying [48].

Using the phase-vocoder digital audio technique, CLEESE first takes the Short-Time Fourier Transform (STFT) of audio files, which decomposes each frame (segment) of the audio file into its frequency coefficients. Then CLEESE modifies the frames' STFT coefficients as required. For example, it shifts a frame's frequency coefficients to higher frequency positions to achieve a higher pitch [48]. After applying the modifications, CLEESE then generates a modified time-domain signal from the manipulated frames by applying a variety of techniques to ensure continuity or phase-coherence of the resulting sinusoidal components [48].

CLEESE operates by passing user-defined or random breakpoint functions (BPFs) to a spectral processing engine together with other parameters for processing of the sound. The BPFs are functions that determine how transformations vary over the duration of the sound, in other words, they define one or more segments (time-windows) of the input sound where specified modifications should be applied. For

each BPF, a transformed version of the input sound is generated.

For the pitch, time and gain transformations, the BPFs are temporal and are specified as two-column matrices. Each row (breakpoint) in a BPF matrix has two elements: time and value. The time indicates where the next modification should begin from, and the value indicates the amount of modification to be applied. The desired transformation is specified separately in a configuration file. With the specified transformation, CLEESE modifies the input sound along the corresponding dimension (pitch, time, or amplitude) while maintaining the other dimensions constant. CLEESE can also perform chained transformations; for example, apply pitch shifting followed by time shifting.

B. CLEESE TRANSFORMATIONS

1) PITCH-SHIFT TRANSFORMATION

Pitch-shifting involves shifting or displacing the fundamental frequency in a given audio frame to a different (higher/lower) frequency. In this study, the fundamental frequencies are shifted to higher frequency points specified in the BPFs along with the corresponding times where the modifications should start. To determine the new frequency point, CLEESE takes a pitch-shift factor, a value expressed in units of cents (a cent is one hundredth of a semitone), provided in the BPF and uses it to compute the new frequency with respect to the original frequency. As an example, to shift the pitch of the input audio by 2 semitones, a pitch-shift factor of 200 cents is provided in the BPF. Pitch-shift factors less than 0 cents correspond to lowering the pitch, factors greater than 0 cents correspond to raising the pitch, and a factor of 0 cents implies no change or shift in pitch [48].

2) TIME-STRETCH TRANSFORMATION

The time-stretching transformation involves shifting the audio frames from their original positions to earlier or later points. Similarly, for the time-stretching, CLEESE takes a time-shift factor from the given BPF and uses that to determine the new position of a frame. A time-shift factor less than 1 corresponds to compressing the sound, a factor greater than 1 corresponds to stretching the sound, and a factor of 1 implies no change in the original sound duration [48]. For example, using a time-shift factor of 2 doubles the duration of the audio, i.e., a 3-second-long audio will become 6-seconds-long after modification, if the modification is applied to the full length of the input audio.

C. WAV2VEC2 FORCED ALIGNMENT SYSTEM

The wav2vec2.0 forced alignment system^{1,2} uses the wav2vec2.0 [4] ASR model for extracting acoustic features from the audio and estimating the frame-wise label probabilities. It then constructs a Trellis matrix using the ground-truth

¹https://github.com/pytorch/audio/blob/main/examples/tutorials/forced_alignment_tutorial.py

²https://pytorch.org/audio/stable/tutorials/forced_alignment_tutorial.html

TABLE 2. The wav2vec2.0 alignment system outputs all words in an utterance, their respective start and stop times, as well as the confidence score for each alignment.

Confidence_level	Word_label	Start_time	Stop_time
0.53	SHE	0.604	0.725
0.80	HAD	0.765	0.926
1.00	A	0.967	0.987
0.84	THIN	1.108	1.430
0.80	AWKWARD	1.792	2.175
0.91	FIGURE	2.236	2.538

transcript of the utterance, which shows the probability of the transcript’s labels at each timestep. The system then finds the most likely path through the Trellis matrix, producing the alignments between the ground-truth transcript’s words and the spoken audio. The output of the forced alignment process is the start and end timestamps for all words in an utterance as shown in table 2.

D. SPEAKER EMBEDDINGS

A speaker embedding is simply a representation of a speaker’s identity in the form of a fixed size vector given an utterance, and regardless of the utterance duration. Speaker embeddings can be plotted in an embedding vector space to visualize how multiple speakers relate to each other. Speaker embeddings are commonly used for speaker recognition tasks [16], [49] and more recently, to improve multi-speaker TTS models [8]. In addition to the speaker identities, speaker embeddings may carry information about other paralinguistic information such as prosody or emotion and gender of a speaker.

Different approaches have been proposed to encode speaker embeddings, and these include identity vectors (i-vectors) [50], which are low-dimensional projections of the differences between a speaker’s pronunciations and the respective overall average pronunciations; (d-vectors) [51], which are deep neural network (DNN) based and extracted from a hidden layer of a model trained to predict speaker identities; and x-vectors [52], which are also DNN based but capture segment/utterance level information as well as frame-level information by using either statistical or max-pooling method to gather the frame level information as segment level representation [53].

III. METHODOLOGY

In this section, we describe the implementation of the proposed adult-to-child speech augmentation process. The python toolkit, CLEESE, is used to perform two key transformations to the adult speech data with the aim of transforming them to child-like speech. Fig. 1 shows a flow diagram of the overall augmentation process.

First, we triage the adult speakers by comparing the cosine similarities of their speaker embeddings to child speaker embedding prior to the augmentation process, see Fig. 2. This is done by computing the mean child speaker embedding as

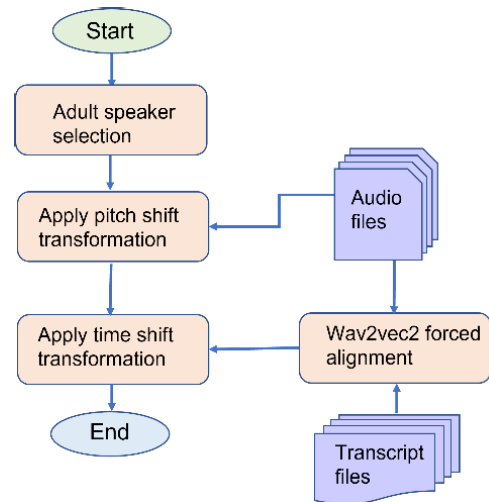


FIGURE 1. Flow diagram for the adult-child speech augmentation process.

well as the mean speaker embedding per adult speaker. Each adult speaker’s cosine similarity to the mean child embedding is computed, and the value is compared to a threshold value for a selection decision to be made. More details on this in section VI, B.

Next, we apply the pitch-shifting transformation to the utterances of the selected speakers. For each utterance, the pitch transformation is applied to the full utterance length. To achieve this, a BPF is created with one breakpoint (time: start of utterance and value: the desired pitch-shift factor e.g. 100 cents i.e., 1 semitone). CLEESE applies the transformation from the specified time stamp to the end of the utterance unless another breakpoint is encountered. Therefore, a single breakpoint (row) in the BPF modifies the full utterance.

Next, the time-stretching transformation is applied to the pitch-shifted utterances. To successfully stretch the desired segments of the sound, the exact start and stop times for the segments are needed to create the appropriate BPFs for stretching. For this, the wav2vec2.0 based forced alignment system is employed to align the adult speech with their corresponding transcripts. Based on the word timestamps, the start and end times of all “white spaces” in the utterance are derived and used in creating BPFs for the time-stretching transformation. The start time for each word and white space is used as a breakpoint in the BPFs, and different stretch factors are used for words vs whitespaces.

IV. EXPERIMENTS

The proposed techniques for augmentation were implemented on an NVIDIA GeForce RTX 2080 Super GPU, and to scale our experiments we used an NVIDIA RTX A6000 GPU.

A. PRELIMINARY TESTS

Pitch-shift and time-stretch transformations were applied to randomly selected subsets of two adult speech datasets: LJ

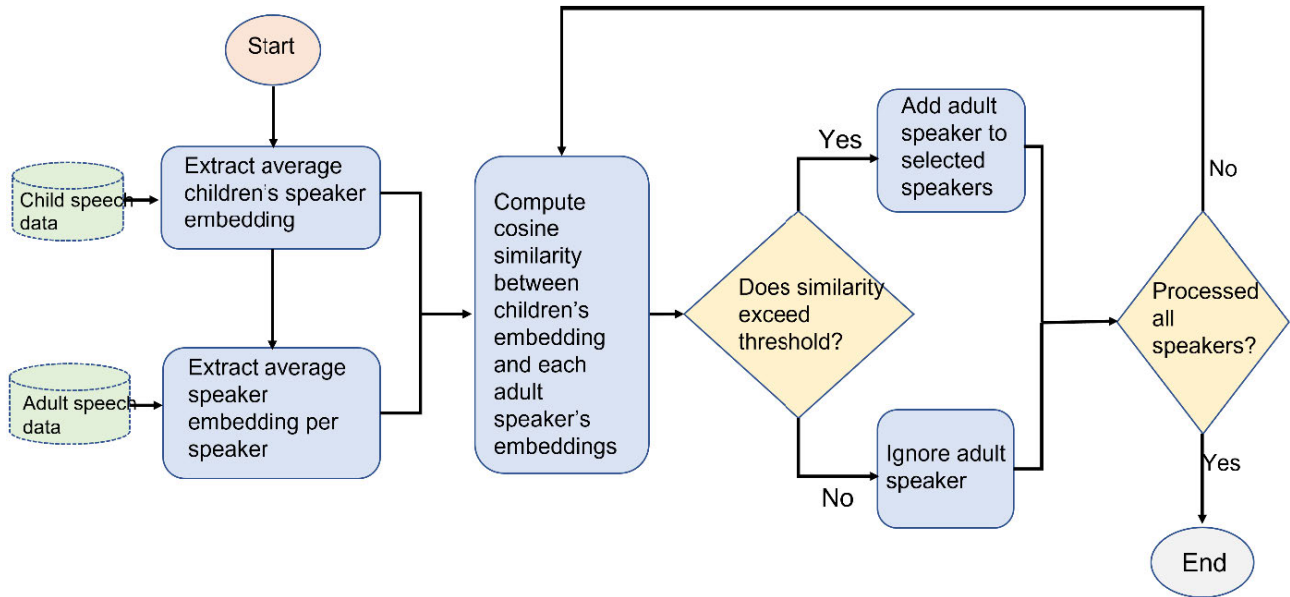


FIGURE 2. A flow diagram for the adult speaker selection process.

speech [31] and voxceleb1 [34]. Pitch-shift factors in the range of 100 cents (1 semitone) to 800 cents (8 semitones) were tested on both male and female speakers, and random time-shift factors in the range of 2 to 4 were also tested. The goal was to determine approximately the range of pitch-shift and time-shift factors that will make sense to use in future experiments.

The time-stretched utterances were qualitatively evaluated by listening to them, and it was observed that a time-shift factor of 4, which quadruples the audio length, resulted in extremely sluggish augmented utterances, and even a factor of 2, which doubles audio length, resulted in utterances that were still a bit too slow. Another observation made was that stretching the individual words in the utterance made them sound unrealistic.

For the pitch-shift transformation, we observed that, with the range of values (pitch-shift factors) that achieved desired results on some of the speaker identities, other speaker identities did not sound realistic, even after extending the range of pitch-shift factors. From these initial tests we determined that not all adult voices can be successfully tuned to sound child-like.

To resolve this and allow a larger study to be conducted, it was necessary to first triage and determine the adult speakers whose voices are more suitable for transforming into natural child voices. This could be achieved by projecting both adults' and children's speaker embeddings into a latent speaker embedding space for comparison.

B. INITIAL EXPERIMENTS

1) COMPARISON OF ADULTS' AND CHILDREN'S SPEAKER EMBEDDINGS

To compare the adults' and children's speaker identities, a Generalized End-to-End (GE2E) Loss [49] based speaker

embedding (encoder) tool known as Resemblyzer [54] was used. It uses a d-vectors based speaker encoder model [49] which uses the GE2E loss for optimization. It also has multiple functionalities for visualizing and comparing the extracted embeddings using the Unified Manifold Approximation and Projection (UMAP) for dimension reduction.

Initially, the speaker embeddings of multiple speakers (both adult and children) were plotted via UMAP with the aim of finding adult speaker embeddings closest to the children's speaker embeddings. In Fig. 3, we show some speaker embeddings in a UMAP plot for visualization.

All male speaker embeddings are marked with black crosses, female speaker embeddings are marked with blue triangles and all child speaker embeddings are marked with red circles. The children's embeddings cluster in a small section of the embedding space. However, it was challenging to accurately identify the adult speakers that are closest or most similar to children by visual inspection. Therefore, it was decided to perform a cosine similarity-based comparison and select speakers with the highest similarity values for the main augmentation experiments.

2) COMPARISON OF EMBEDDINGS BASED ON COSINE SIMILARITY

The cosine similarity score is a number between 0 and 1. A similarity of 1 means the two embeddings compared are identical, and a similarity of 0 means they are completely different. Firstly, we extracted the speaker embeddings for multiple child speakers taken from the CMU kids corpus [29]. From previous research [24] as well as initial experiments (see Fig. 3), it is known that children's speaker embeddings form a small cluster in the speaker embedding latent space;

Embedding Projections

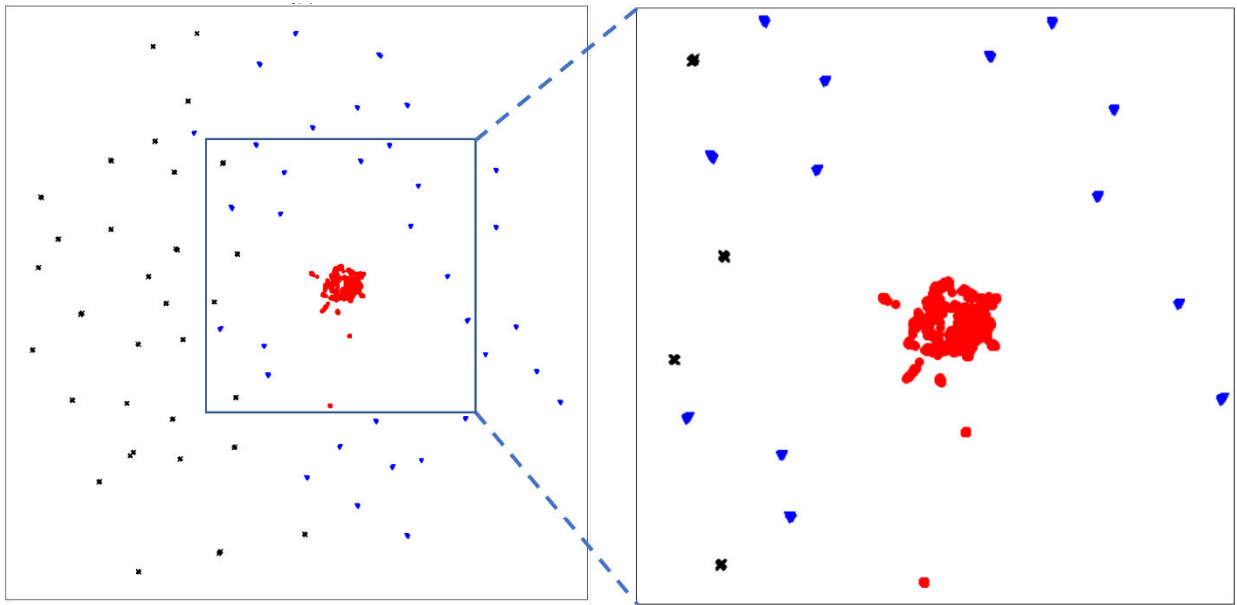


FIGURE 3. Projection of 65 Adult speaker embeddings from Librispeech: 31 male (black), 34 females (blue) and 31 child speaker embeddings from CMU Kids.

hence, we computed the mean child speaker embedding to represent the children embedding cluster.

Next, we took a random subset from the train-clean-100 subset of the Librispeech dataset [32] and computed an average speaker embedding for each adult speaker, by averaging the embeddings of their individual utterances. Then we compared them to the mean child speaker embedding using the cosine similarity metric. A flow diagram showing the flow of this process is seen in Fig. 2.

All adult speakers whose embeddings exceeded a pre-defined threshold of 0.65 were selected for augmentation as in equation 1. This threshold was chosen by listening to some of the utterances and observing their corresponding similarities. Fig. 4 shows some examples of the cosine similarities computed. More statistics regarding the cosine similarities are shown in the next section.

$$Dec = \begin{cases} 0, & sim_score < 0.65 \\ 1, & sim_score \geq 0.65 \end{cases} \quad (1)$$

where Dec is adult speaker selection decision, sim_score is the computed cosine similarity score between an adult’s speaker embedding and the average child speaker embedding.

3) AUGMENTATION PROCESS

Further tests were done on the selected speakers (i.e., adult speakers whose cosine similarities exceeded the threshold) thereafter. Two separate ranges of pitch-shift factors were empirically chosen for the two genders. This was done by listening to the pitch-shifted utterances and rating them in terms of how convincingly child-like they sounded. For male

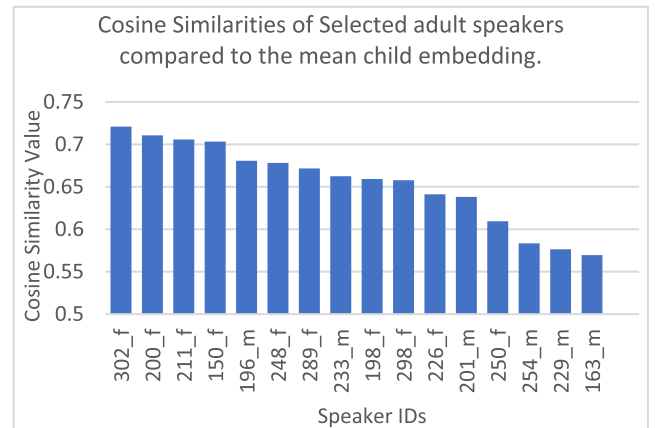


FIGURE 4. Original cosine similarities showing how similar Librispeech adult speakers are to mean CMU kids child speaker embedding. The speaker gender is suffixed to the Speaker IDs.

and female speakers, the ranges of 500 to 700 cents and 100 to 300 cents were chosen respectively.

Based on the observations made about the time-stretched utterances in the preliminary tests, it was decided to stretch only the pauses (whitespaces) between all words in the utterances, as well as the unusually long words, without stretching every single word. For stretching all the whitespaces, we first used a time-stretch factor of 2 (i.e., we doubled the length/duration of pauses) and then reduced it to a time-stretch factor of 1.8 after qualitatively evaluating a few of the augmented utterances. In addition, we identified the unusually longer words in the utterances - that might be

TABLE 3. Number of librispeech train-clean-100 speakers above and below the cosine similarity threshold.

Speakers	Total No.	No. speakers above threshold	No. speakers below threshold
All	251	121	130
Males	126	27	99
Females	125	94	31

TABLE 4. Final shift factors used for pitch and time transformations.

Transformation	Shift-factors
Pitch-shift (female)	100, 200, 250, 300, 350, 400 (cents)
Pitch-shift (male)	500, 600, 700 (cents)
Time-stretch	1.8 (for white spaces), 2.0 (for longer words)

difficult for children to pronounce - and stretched them using a factor of 2. This was done by computing the duration of each word and comparing it to an empirically chosen word length threshold.

C. MAIN EXPERIMENTS

For the main experiments, we used all the data in the train-clean-100 subset of Librispeech [32] as the adult speech dataset. A subset of the CMU kids dataset [29] was used as the child speaker set, specifically the Fort Pitt (FP) subset. Firstly, the adult speakers most proximate/similar to children in terms of speech/voice were determined by performing the cosine similarity comparison explained in Section III, B using the same decision threshold value of 0.65 as in the initial experiment. Table 3 shows the number of adult speakers above and below the cosine similarity threshold.

Once the most similar speakers were selected, the two augmentation techniques explained in Section II, B; namely, pitch-shifting and time-stretching transformations were applied to all individual utterances of the selected speakers. The same pitch-shift and time-stretch factors chosen in the initial experiment were applied here. First, we applied the pitch-shifting transformations and then applied the time-stretching transformation on the output of the pitch-shifting transformation. Table 4 shows the final shift factors used for the pitch-shift and time-stretch transformations.

This resulted in multiple sets/folders of data per speaker, each containing utterances augmented with different augmentation parameters. Specifically, the sets of utterances differed in terms of pitch-shift factors only, as the time-stretching parameters were kept constant for all sets and all genders.

D. OBJECTIVE EVALUATION

In the initial experiment section, the cosine similarity value served as a good metric to determine the proximity of adult speaker embeddings to the average child speaker embedding. Therefore, to objectively evaluate the augmented speech,

TABLE 5. Statistics of cosine similarities for librispeech train-clean-100 before and after augmentations.

Data	Before augmentation		After augmentation	
	Mean	STD	Mean	STD
All above threshold (121)	0.69	0.032	0.83	0.034
20 spkrs in MOS study	0.740	0.022	0.86	0.027
16 spkrs in MOS	0.745	0.019	0.87	0.018

it made sense to recompute the cosine similarities between adult speakers' average embedding (after augmentation) and the average child embedding. After recomputing the cosine similarities, we observed that there was a general increase in the similarity values for all the speakers. Fig. 5 below shows the cosine similarities of selected speakers before and after transformations were applied. A similarity score of 1 would indicate that a speaker is exactly the same as the average child speaker. Table 5 also shows statistical analysis of the adult speech data before and after augmentation. Note that the cosine similarities of all individual child speakers' embeddings to the mean child embedding were in the range of 0.9 to 0.973, except one child (0.837).

E. SUBJECTIVE EVALUATIONS

While the increase in the cosine similarity of an augmented adult speaker gives a strong indication that the augmentation pipeline is achieving its primary goal, it is not possible to judge how realistic or intelligible the augmented voice is. In the case of some subjects, it was noted that while the cosine similarity was high, the corresponding speech was occasionally distorted and unrealistic.

For this reason, it was decided to conduct a human listener evaluation study to validate how realistic the augmented speech from a speaker is and confirm that it remains intelligible. Such a study can also help confirm the best speakers and the optimal augmentation parameters to use for individual speakers to build a larger augmented speech dataset – a core goal of this research.

To subjectively evaluate the augmented speech samples, the MOS [55] subjective evaluation method was applied. MOS evaluation is widely used to evaluate speech models, such as TTS and Voice Conversion (VC) models, by asking human evaluators to rate various aspects of speech quality such as naturalness, intelligibility, similarity, etc.

1) DESIGN OF MOS STUDY

There were three specific goals for the study: i) Determine the optimal pitch-shift factor per speaker, ii) Determine how realistic (convincingly child-like) the augmented utterances sound and iii) Determine whether the augmented utterances are distorted beyond understanding or if they remain intelligible. To achieve the goals of the study, three questions that

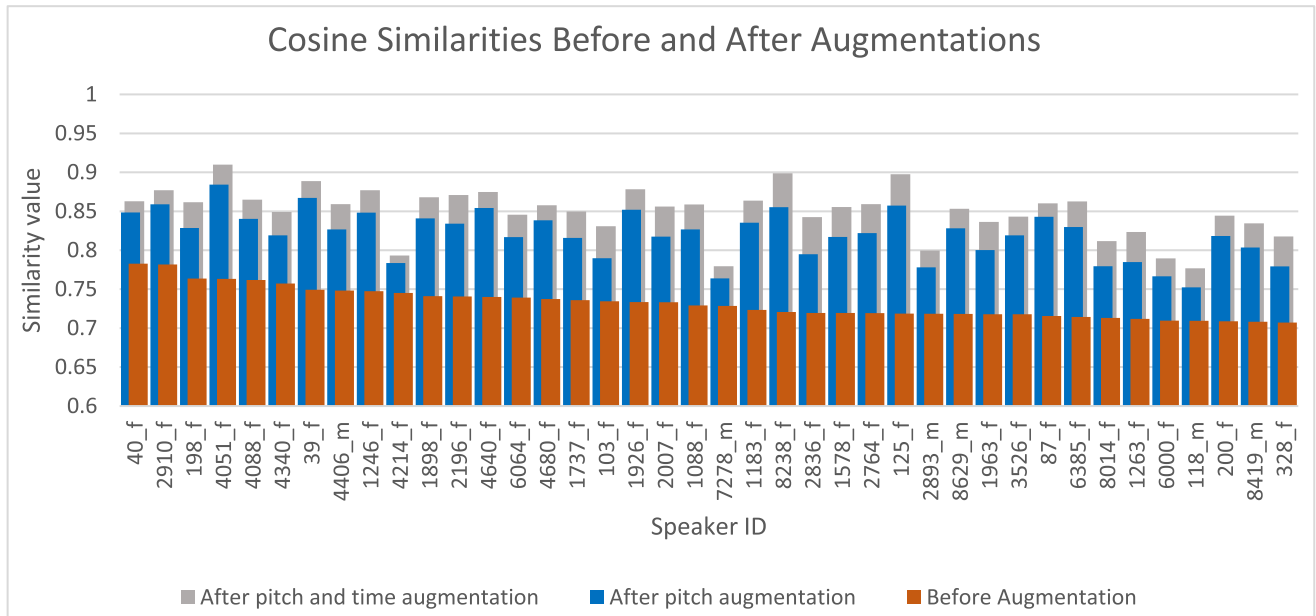


FIGURE 5. Increases in cosine similarity between adult and child speaker embeddings after pitch shifting and time stretching.

capture the information required were chosen and presented to the evaluators.

For the first goal, evaluators were provided with multiple variations per utterance and asked to select the most child-like sounding one. The difference between the variations are the pitch-shift factors used in the pitch-shift transformation. The sample selected for the first question is used in the remaining questions. Secondly, evaluators were asked to rate the selected sample in terms of how convincingly child-like or how realistic it sounds on a scale of 1 to 5. Note that the linguistic contents of the utterances are adult-like and very different from the typical linguistic content of child speech. Evaluators were given prior notice and were asked to disregard the adult-like linguistic content while rating the convincingness.

Thirdly, evaluators were asked to rate the same selected sample in terms of intelligibility on a scale of 1 to 5. Evaluators were restricted to only 5 grading points (i.e., 1, 2, 3, 4 and 5); they were not allowed to give intermediate scores, such as 2.5. Evaluators were also asked to identify the gender of the speaker by choosing one of three options: Boy, Girl and Can't say. Finally, evaluators were also given the option to leave comments if they had any. Table 6 shows explanations of the scales for convincingness (question 2) and intelligibility (question 3) following approach in [24].

With this design, a first MOS study (Study A) was conducted on utterances augmented with the following pitch-shift factors: 100 cents, 200 cents and 300 cents, meaning for each utterance, there were three variations for evaluators to choose from. After this first study was completed and the results were processed, it was decided to conduct a second MOS study (Study B) to refine the outcome of the

first. In particular, we wanted to see the effect of including utterances augmented with higher and more finely granulated pitch-shift factors as compared to the first study: 250, 300, 350 and 400 cents. This is because in the first study, the same variation of utterance (augmented with the highest pitch-shift factor of 300 cents) was selected by almost every evaluator as the most child-like for all female speakers, so the range of pitch-shift factors to investigate clearly needed expansion.

Study A:

In the first evaluation, there was a total of 30 evaluators, mainly drawn from an undergraduate engineering class. These were divided into two groups of 15 evaluators. Augmented speech samples were taken from Librispeech speakers. 20 speakers were chosen for the MOS study, after listening to samples from all their recording sessions to check for noise and rate the quality. This was done after triaging the adult speakers as described in Section III B. They included 16 female speakers and 4 male speakers with the highest cosine similarities and high-quality audio samples. Each group of 15 evaluators was given a unique set of 10 different speakers to review (8 females and 2 males). The purpose was to reduce the total number utterances per evaluator. To diversify the phrases, each evaluator group was further divided into 3 subgroups and each subgroup was given a unique (randomly selected) set of 2 phrases per speaker, resulting in a total of 20 phrases to evaluate per evaluator. They were given three augmented samples (variations) per phrase: A, B and C corresponding to pitch shifting factors of 100, 200 and 300 cents, respectively.

Study B:

In the second evaluation study, augmented samples from only the 16 female speakers out of the top 20 speakers

TABLE 6. Explanation of rating scales for question 2 (convincingness) and 3 (intelligibility).

Rating (Score)	Question 2: Convincingness	Question 3: Intelligibility
1	Unconvincing	Bad voice intelligibility and bad word comprehensibility; None/only a few words are intelligible.
2	Slightly convincing, not very like a normal child's voice	Weak voice intelligibility and weak word comprehensibility; more than 50% of the words unintelligible
3	Plausible but not very convincing as a normal child's voice	Weak voice intelligibility but plausible word comprehensibility; more than 50% of words are intelligible but require significant concentration.
4	Quite convincing, sounds very close to a child's voice	Mediocre but plausible intelligibility and comprehensibility; most words intelligible and relatively easy to identify.
5	Very convincing, sounds exactly like a child's voice	Good intelligibility and comprehensibility; All words are quite clear and comprehensible.

TABLE 7. Summary of the data distribution for mos studies a and b.

	Study A		Study B	
	I	II	I	II
Total no. of Evaluators	30		60	
No. of Evaluators	15	15	30	30
No. of Male IDs	2	2	-	-
No. of Female IDs	8	8	8	8

(the same speakers as in the first study) were evaluated. There were 60 evaluators, again mostly engineering students, divided into 2 main groups of 30 evaluators. Each group was further divided into 3 subgroups of 10 students, similar to the approach used in Study A. This time, each evaluator received 16 phrases from 8 speakers: two phrases per speaker as in the first evaluation. Specifically, there were four variations per utterance/phrase: A, B, C and D corresponding to pitch shifting factors of 250, 300, 350 and 400 cents, respectively. Information about evaluators for the two MOS studies is presented in Table 7.

V. RESULTS AND DISCUSSION

In Section III, we described our adult-to-child speech augmentation experiments using the two augmentation techniques described in Section II, with a goal of making the adult voices sound child-like. We also conducted two MOS studies to evaluate the quality of the synthetic child-like speech. In this section, we present and discuss the results of our experiments.

Tables 8 and 9 show the results obtained from the first and second MOS studies, respectively. More detailed presentations of the MOS evaluation results are shown in Tables 12 and 13 in the Appendix. The results of the subjective evaluation showed that the utterances of adult female speakers consistently ranked with higher scores for intelligibility and significantly higher scores for

TABLE 8. Mean and standard deviation (std) of convincingness and intelligibility MOS scores (C-MOS and I-MOS) from Study A.

Speakers	No.	C-MOS (STD)	I-MOS (STD)
Female	16	3.37 (0.37)	4.32 (0.20)
Male	4	1.76 (0.37)	3.87 (0.39)
All	20	3.05 (0.75)	4.23 (0.30)

TABLE 9. Mean and standard deviation (std) of convincingness and intelligibility MOS Scores (C-MOS and I-MOS) from Study B.

Speakers	No.	C-MOS (STD)	I-MOS (STD)
Females	16	3.70 (0.35)	4.36 (0.25)

TABLE 10. Details of the synthetic and original data used in finetuning.

Finetuning data	Adult/child	Orig/Synt	Duration(hrs)
Original_12h	Adult	Orig	12
Augmented_17h	Child	Synt	17
Original_220h	Adult	Orig	220
Augmented_311h	Child	Synt	311
MyST_55h	Child	Orig	55

convincingness. We had anticipated this result as only 4 males ranked in the top 20 speakers from Librispeech train-clean-100. It is clear that female speakers offer a better starting point to build synthetic child voices than male speakers.

As shown in both Table 12 and Table 13, the optimal pitch-shifting factors for the female speakers lie in the range of 300 to 400 cents. Augmenting the pitch above this range causes the augmented speech to sound more chipmunk-like rather than child-like. For the male speakers, the pitch-shift factor of 600 cents was selected for 3 out of 4 speakers but the augmented speech were unconvincing as child voices, with a very low average MOS score of 1.76.

TABLE 11. WER of ASR models finetuned with synthetic and original speech data.

Model	Group	Pretraining	Finetuning dataset	WER	WER	WER	WER
				MyST_10h	PFS_10h	CMU_9h	Devclean_9h
1	Group A	Librispeech	Original_12h	19.95	25.10	18.95	5.78
2		Librispeech	Augmented_17h	20.11	20.48	19.14	6.58
3		Librispeech	Original_12h + Augmented_17h	18.17	18.11	16.07	5.49
4	Group B	Librispeech	MyST_55h	8.13	17.67	16.47	7.72
5		Librispeech	MyST_55+Original_12h	8.10	16.76	15.45	5.62
6		Librispeech	MyST_55h + Augmented_17h	7.98	14.015	15.02	4.87
7		Librispeech	MyST_55h +Original_12h+Augmented_17h	7.95	14.85	13.92	5.54
8	Group C	Librispeech	Original_220h	15.09	16.59	14.41	4.39
9		Librispeech	Augmented_311h	17.42	15.86	15.09	4.83

The overall C-MOS score of the most child-like samples was approximately 3.0 when adult male speakers were considered, and 3.7 when only adult female speakers were evaluated (study B). Both convincingness MOS values are above average and implies that the augmented samples are reasonably convincing in terms of human perception and very convincing when only female speakers are used in the study.

A relatively higher I-MOS was obtained for the augmented samples of both genders, showing that generating synthetic child voices using our proposed method does not significantly degrade the intelligibility of the original speech samples.

Note that there are limitations in going from adult speech to child speech; for example, the linguistic content of adult speech data is completely different from the typical linguistic content of children's speech. For this reason, tuning the pitch and speaking rate of adult speech would not make the speech sound completely natural as child speech in terms of the linguistic content. However, these tunings can make the voices alone sound reasonably child-like, which is the target for the current study.

The mean cosine similarity of adult speakers after augmentation was 0.83 for all speakers exceeding the similarity threshold and 0.87 for the top 16 female speakers (see Table 4), whereas the mean cosine similarity of the individual child speakers was 0.94, indicating that there is still potential to further augment the adult speakers to sound closer to child speakers. This suggests that additional prosodic features and paralinguistic elements could be investigated and added into our augmentation strategy to improve the cosine similarity score of the adult speakers.

Finally, to validate the augmented child speech data in a practical application, we next run some ASR fine-tuning experiments, as presented in the next section.

VI. VALIDATION OF THE AUGMENTED SPEECH: EXAMPLE APPLICATION – ASR FINETUNING

In this section, as an example application, we conduct semi-supervised ASR finetuning experiments with our augmented

adult speech dataset, to show that the augmented speech could achieve improvement over simply using additional adult speech to finetune the ASR for child speech.

Note that the main goal of our study was to explore data augmentations to make adult speech data sound more child-like (i.e. closer to child speech data) in order to provide more child-like data for training, testing and validation of ASR and TTS models to improve their performance on real child speech. Here, we show that finetuning a semi-supervised ASR model with augmented adult speech data can improve the ASR model's performance on child speech. We show that even when finetuned with adult-only speech data, the performance of the model improves to an extent; however, there is some additional improvement when the augmented adult speech is used.

We used the state-of-the-art (SOTA) wav2vec2.0 ASR model [3], which uses a self-supervised learning approach and has a two-step training process. First, the model is pretrained on a large amount on unlabeled speech, then it is finetuned on labelled speech data for a downstream task, such as ASR. We used a publicly available pretrained wav2vec2.0 model, which was trained on approximately 1000 hours of unlabeled Librispeech data [32]. This model was then finetuned with different combinations of our augmented datasets in the various finetuning experiments as presented in the next sub-section. The aim was to compare the performance of an ASR model finetuned with real child and/or adult speech versus the same model finetuned with our augmented data (synthetic child-like speech). The Word Error Rate (WER) metric was used to measure the performance of the finetuned ASR models.

A. ASR FINETUNING DATASETS

We created two sets of synthetic child speech:

- *Augmented_17h*: Contains augmented utterances from the 16 female speakers of the train-clean-100 Librispeech dataset, whose speaker embeddings are most similar to an average child embedding from the

CMU-kids corpus by cosine similarity. The female speakers were selected by ranking all female speakers by their similarity score. This data totals approximately 17 hours in duration.

- *Augmented_311h*: Contains augmented utterances of all female speakers in Librispeech train-clean-360, train-clean-100, dev and test sets combined, whose similarity score to the average real child embedding from the CMU-kids corpus is above 0.6. This data totals approximately 311 hours in duration.

We also used original (non-augmented) adult speech from Librispeech [32] and real child speech data from the MyST child speech corpus [30] for our finetuning experiments:

- *Original_12h*: Contains 12 hours of original adult speech.
- *Original_220h*: Contains 220 hours of original adult speech.
- *MyST_55h*: Contains 55 hours of cleaned MyST child speech, which was prepared according to [56]

The *Original_12h* and *Original_220h* sets are the original Librispeech (adult speech) counterparts of the *Augmented_17h* and *Augmented_311h* sets, respectively. Note that there is an increase in the number of hours of speech data when augmenting from *Original_12h* to *Augmented_17h* and from *Original_220h* to *Augmented_311h*. More information about the finetuning datasets can be found in Table 10.

B. ASR FINETUNING EXPERIMENTS

To test our hypothesis of a lower WER on child test data after finetuning on our synthetic child-like speech data, we prepared multiple finetuning experiments. The details of these experiments are presented in Table 11. The experiments were divided into three groups- A, B and C. Group-A experiments contained only the Original and Augmented datasets. MyST_55h was added for the finetuning experiments in Group-B in addition to the Original and Augmented datasets. Group-C experiments used the combined Librispeech datasets across all speakers, both original and augmented versions. All the groups used a pretrained wav2vec2.0 model which was pretrained on 960 hours of Librispeech data.

We used four test datasets to test our finetuned models at the inference stage. These datasets were prepared in accordance with our previous research on child speech ASR [56]. Since MyST [30] is the largest child audio corpus available publicly for research use, it was used for both finetuning and inference. This was done to see the performance when finetuning and testing on similar data distributions. We used 10 hours of MyST child speech data, 10 hours of PFSTAR British English data [27], 9 hours of CMU-Kids American English child speech data [29], and 9 hours of Librispeech dev-clean data as our test datasets. Different child speech test datasets were selected specifically to check the performance of our finetuned models on datasets that have different acoustic

attributes, in conjunction with adult speech also. WER values obtained on these test datasets during inference are shown in Table 11.

C. ASR FINETUNING RESULTS

Group-A: Finetuning with *Augmented_17h* resulted in a decrease in WER on the PFS_10h data (British English child speech), and a slight increase in WER on the other child test sets, when compared to inference with a model finetuned on its original speech counterpart (*Original_12h*). Furthermore, combining just 17 hours of the augmented child speech (*Augmented_17h*) with original adult speech (*Original_12h*) leads to a slight improvement in WER on all child test sets, as well as the adult speech test data.

Group-B: This group uses the cleaned MyST_55h dataset in addition to the datasets used in Group-A experiments. Using Augmented data along with MyST child speech dataset led to a decrease in WER on all the test datasets (see model 6 in Table 10).

Group-C: This group used datasets created from large-scale augmentation. There was an 18.3x increase in dataset size from *Original_12h* to *Original_220h* and from *Augmented_17h* to *Augmented_311h*, respectively. Augmentation led to a decrease in WER on PFS_10h test data, but an increase in WER for all other datasets, which is very similar to the results of Group-A experiments.

D. DISCUSSION OF RESULTS

For Group-A, the WER decreases for all the test datasets when both original and augmented adult speech datasets were used for finetuning.

With MyST data inclusion in Group-B, we see a major decrease in WER compared to Group-A results.

Furthermore, in Group-B, it can be seen that adding augmented speech along with MyST_55h (model 6) led to decrease in WER on all the test datasets compared to using only MyST child speech for finetuning (model 4) or using both MyST and original adult speech (model 5). Also, by adding both the original and augmented speech for finetuning (model 7), an increase in WER can be observed on PFS_10h and adult data, while the WER on CMU_9h is reduced.

Using *Original_220h* and *Augmented_311h* in the Group-C experiments did not lead to improvements in ASR performance when compared with Group-B results. Comparing models 2 and 9, with an 18x increase in the amount of augmented data, respectively, the WER decreased by only 3.5 points on average on child speech.

While improvements in the child ASR performance were expected, the results from the example application do not show significant improvements using just the large amount of synthetic child speech for finetuning. This could partly be attributed to a lack of natural prosody in the augmented adult data (synthetic) when compared to real child audio. Although the synthetic speech sounds reasonably child-like in

TABLE 12. Per speaker MOS Scores and best shift factors from 1ST evaluation.

Speaker ID	True Gender	Shift Factor Selection Count			Best Shift Factor	Convincingness	Intelligibility
		100 (A)	200 (B)	300 (C)			
39	Female	1	2	27	C	3.37	4.27
40	Female	3	9	18	C	3.4	4.3
103	Female	4	6	20	C	3.87	4.57
198	Female	1	3	26	C	3.2	4.07
1183	Female	0	6	24	C	3.13	3.87
1624	Male	2	18	10	B	2.13	4.13
1737	Female	1	4	25	C	3.77	4.43
1898	Female	0	4	26	C	3.13	4.27
1926	Female	1	3	26	C	3.8	4.2
2893	Male	12	9	9	A	2.03	4.27
2007	Female	0	4	22	C	3.31	4.42
2196	Female	0	0	26	C	2.42	4.31
2764	Female	0	6	20	C	3.31	4.46
2910	Female	4	5	17	C	3.46	4.04
4051	Female	0	2	24	C	3.92	4.5
4340	Female	1	8	17	C	3.46	4.31
4680	Female	3	1	22	C	3.04	4.54
1088	Female	0	5	21	C	3.31	4.54
4406	Male	9	15	2	B	1.46	3.42
8419	Male	8	11	7	B	1.42	3.65

TABLE 13. Per speaker mos scores and best shift factors from 2ND evaluation.

Speaker ID	True Gender	Shift Factor Selection Count				Best Shift Factor	Convincingness	Intelligibility
		250 (A)	300 (B)	350(C)	400 (D)			
39	Female	2	9	16	17	D	4.02	4.71
40	Female	10	12	14	8	C	3.75	4.61
103	Female	10	10	14	10	C	3.82	4.5
198	Female	2	8	16	18	D	3.82	4
1183	Female	9	14	10	11	B	3.11	3.95
1737	Female	4	16	11	13	B	3.79	4.34
1898	Female	5	11	19	9	C	3.20	4.45
1926	Female	4	8	19	13	C	4.16	4.61
2007	Female	4	12	13	11	C	3.85	4.4
2196	Female	7	8	13	12	C	3.4	4.675
2764	Female	7	12	6	15	D	3.63	4.13
2910	Female	7	8	13	12	C	3.78	3.9
4051	Female	8	9	14	9	C	3.98	4.35
4340	Female	9	12	9	10	B	4.1	4.43
4680	Female	6	9	11	14	D	3.0	4.3
1088	Female	8	10	10	12	D	3.83	4.48

terms of pitch and speaking rate, they are still lacking natural prosody characteristics such as stammering, long pauses (due to uncertainty) and other features seen in real child audio

recordings. Features of natural child speech prosody could be modeled in addition to the proposed augmentation approach, which is expected to improve WER further.

VII. CONCLUSION AND FUTURE WORK

We have presented experiments exploring the possibility of generating synthetic child voices by augmenting existing adult speech datasets. Augmenting the pitch and duration of adult speech samples generally caused them to sound more child-like, however this worked better for the female adult speakers as compared to the males. This observation was further confirmed by the results of a subjective evaluation conducted using the MOS evaluation method with a total of 72 participants. The average cosine similarity of augmented adult speech is still lower than that of real child speakers, therefore more research is required to improve the similarity of augmented speech. While the improvements in the performance of finetuned ASR models on real child speech are relatively small, they provide a validation of the approach which can be further improved with a more sophisticated set of augmentations. These are planned for future work.

We have scaled this data augmentation process to provide a large amount of synthetic child speech suitable for training child-friendly TTS, VC and ASR models and the data, along with pipeline implementation code, will be made publicly available to other researchers who wish to replicate our approach.

In future experiments, we plan to investigate and apply other tuning techniques to better augment the adult male voices as well as improve the existing augmentation techniques to better suit the linguistic content on a sentence-by-sentence basis. We also plan to investigate methods that take the natural child prosody or paralinguistic feature modeling into consideration; this could contribute to further increasing the similarity of the augmented adult (synthetic child) speech to real child speech.

APPENDIX

See Tables 12 and 13.

ACKNOWLEDGMENT

The authors would like to thank Zoran Fejzo from Xperi Corporation and the rest of the team members for their helpful discussions and feedback.

REFERENCES

- [1] D. Amodei et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," 2016, *arXiv:1512.02595*.
- [2] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," Aug. 2015, *arXiv:1508.01211*. Accessed: Jan. 12, 2023.
- [3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," Apr. 2019, *arXiv:1904.05862*.
- [4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2020, pp. 12449–12460. Accessed: Jan. 12, 2023. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html>
- [5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," May 2020, *arXiv:2005.08100*. Accessed: Jan. 12, 2023.
- [6] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," Apr. 2019, *arXiv:1904.08779*, doi: [10.21437/Interspeech.2019-2680](https://doi.org/10.21437/Interspeech.2019-2680).
- [7] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6124–6128, doi: [10.1109/ICASSP40776.2020.9053889](https://doi.org/10.1109/ICASSP40776.2020.9053889).
- [8] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," Jan. 2018, *arXiv:1806.04558*. Accessed: Jan. 11, 2023.
- [9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," Apr. 2017, *arXiv:1703.10135*. Accessed: Apr. 7, 2022.
- [10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," Feb. 2018, *arXiv:1712.05884*. Accessed: Jan. 12, 2023.
- [11] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," Mar. 2021, *arXiv:2006.04558*. Accessed: Jun. 21, 2021.
- [12] S. Beliaev and B. Ginsburg, "TalkNet2: Non-autoregressive depth-wise separable convolutional model for speech synthesis with explicit pitch and duration prediction," Jun. 2021, *arXiv:2104.08189*. Accessed: Jan. 11, 2023.
- [13] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," Oct. 2020, *arXiv:2005.11129*. Accessed: Jun. 21, 2021.
- [14] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," Feb. 2021, *arXiv:2006.06873*. Accessed: Jan. 12, 2023.
- [15] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: An end-to-end neural speaker embedding system," May 2017, *arXiv:1705.02304*. Accessed: Nov. 3, 2021.
- [16] N. R. Kologuri, J. Li, V. Lavrukhin, and B. Ginsburg, "SpeakerNet: 1D depth-wise separable convolutional network for text-independent speaker recognition and verification," Oct. 2020, *arXiv:2010.12653*. Accessed: Aug. 03, 2021.
- [17] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," Aug. 2019, *arXiv:1808.00158*. Accessed: Jan. 12, 2023.
- [18] W. Xie, A. Nagrani, J. Son Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," May 2019, *arXiv:1902.10107*. Accessed: Jan. 12, 2023.
- [19] J. A. C. Nunes, D. Macedo, and C. Zanchettin, "AM-MobileNet1D: A portable model for speaker recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8, doi: [10.1109/IJCNN48605.2020.9207519](https://doi.org/10.1109/IJCNN48605.2020.9207519).
- [20] R. Tong, L. Wang, and B. Ma, "Transfer learning for children's speech recognition," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Dec. 2017, pp. 36–39, doi: [10.1109/IALP.2017.8300540](https://doi.org/10.1109/IALP.2017.8300540).
- [21] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "Spectral modification for recognition of children's speech under mismatched conditions," in *Proc. 23rd Nordic Conf. Comput. Linguistics (NoDaLiDa)*, Reykjavik, Iceland: Linköping Univ. Electronic Press, May 2021, pp. 94–100. Accessed: Jan. 12, 2023. [Online]. Available: <https://aclanthology.org/2021.nodalida-main.10>
- [22] E. Booth, J. Carns, C. Kennington, and N. Rafla, "Evaluating and improving child-directed automatic speech recognition," in *Proc. 12th Lang. Resour. Eval. Conf. Marseille, France: European Language Resources Association*, May 2020, pp. 6340–6345. Accessed: Jan. 12, 2023. [Online]. Available: <https://aclanthology.org/2020.lrec.1.778>
- [23] *End-to-End Neural Systems for Automatic Children Speech Recognition: An Empirical Study | Elsevier Enhanced Reader*. Accessed: Jan. 12, 2023. [Online]. Available: <https://reader.elsevier.com/reader/sd/pii/S0885230821000905?token=B1E7CB46937771A433675363A2F64DB62FDBC8EF67723CC34A2D649560F968357D9F186DF9467BF58F35AAECAB1CDB03&originRegion=eu-west-1&originCreation=20230112172622>

- [24] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis," *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: [10.1109/ACCESS.2022.3170836](https://doi.org/10.1109/ACCESS.2022.3170836).
- [25] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jančovič, and M. Carey, "Speaker recognition for children's speech," in *Proc. Interspeech*, 2012, pp. 1836–1839, doi: [10.21437/Interspeech.2012-401](https://doi.org/10.21437/Interspeech.2012-401).
- [26] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, "Children's speaker verification in low and zero resource conditions," *Digit. Signal Process.*, vol. 116, Sep. 2021, Art. no. 103115, doi: [10.1016/j.dsp.2021.103115](https://doi.org/10.1016/j.dsp.2021.103115).
- [27] M. Russell, "The pf-star british english childrens speech corpus," *Speech Ark Limited*, 2006.
- [28] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids² speech corpus and recognizers," in *Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP)*, vol. 4, Oct. 2000, pp. 258–261, doi: [10.21437/ICSLP.2000-800](https://doi.org/10.21437/ICSLP.2000-800).
- [29] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids corpus," *Linguistic Data Consortium*. Philadelphia, PA, USA: Linguistic Data Consortium, 1997.
- [30] S. Pradhan, R. Cole, and W. Ward, "MyST children's conversational speech," *Linguistic Data Consortium*. Philadelphia, PA, USA: Linguistic Data Consortium, 2021.
- [31] *The LJ Speech Dataset*. Accessed: Jan. 12, 2023. [Online]. Available: <https://keithito.com/LJ-Speech-Dataset>
- [32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210, doi: [10.1109/ICASSP.2015.7178964](https://doi.org/10.1109/ICASSP.2015.7178964).
- [33] C. Veaux, J. Yamagishi, and K. MacDonald, "SUPERSEDED—CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Rainbow Passage Which Speakers Read Out Found Int. Dialects English Arch., Apr. 2017. [Online]. Available: <http://web.ku.edu/~idea/readings/rainbow.htm>, doi: [10.7488/ds/1994](https://doi.org/10.7488/ds/1994).
- [34] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, Aug. 2017, pp. 2616–2620, doi: [10.21437/Interspeech.2017-950](https://doi.org/10.21437/Interspeech.2017-950).
- [35] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, Sep. 2018, pp. 1086–1090, doi: [10.21437/Interspeech.2018-1929](https://doi.org/10.21437/Interspeech.2018-1929).
- [36] P. K. O'Neill, V. Lavrukhin, S. Majumdar, V. Noroozi, Y. Zhang, O. Kuchaiev, J. Balam, Y. Dovzhenko, K. Freyberg, M. D. Shulman, B. Ginsburg, S. Watanabe, and G. Kucsko, "SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," Apr. 2021, *arXiv:2104.02014*. Accessed: Jan. 12, 2023.
- [37] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," 2019, *arXiv:1904.02882*. Accessed: May 5, 2023.
- [38] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," Rainbow Passage Which Speakers Read Out Can be Found Int. Dialects English Arch., Nov. 2019. [Online]. Available: <http://web.ku.edu/~idea/readings/rainbow.htm>, doi: [10.7488/ds/2645](https://doi.org/10.7488/ds/2645).
- [39] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. IEEE 9th Workshop Multimedia Signal Process.*, Oct. 2007, pp. 22–25, doi: [10.1109/MMSP.2007.4412809](https://doi.org/10.1109/MMSP.2007.4412809).
- [40] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 1997, pp. 473–476, doi: [10.21437/Eurospeech.1997-161](https://doi.org/10.21437/Eurospeech.1997-161).
- [41] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999.
- [42] R. Mugitani and S. Hiroya, "Development of vocal tract and acoustic features in children," *Acoust. Sci. Technol.*, vol. 33, no. 4, pp. 215–220, 2012, doi: [10.1250/ast.33.215](https://doi.org/10.1250/ast.33.215).
- [43] B. Ahmed, K. J. Ballard, D. Burnham, T. Sirojan, H. Mehmood, D. Estival, E. Baker, F. Cox, J. Arciuli, T. Benders, K. Demuth, B. Kelly, C. Diskin-Holdaway, M. Shahin, V. Sethu, J. Epps, C. B. Lee, and E. Ambikairajah, "AusKidTalk: An auditory-visual corpus of 3-to 12-year-old Australian children's speech," in *Proc. Interspeech*, Aug. 2021, pp. 3680–3684, doi: [10.21437/Interspeech.2021-2000](https://doi.org/10.21437/Interspeech.2021-2000).
- [44] S. Shahnawazuddin, N. Adiga, H. K. Kathania, and B. T. Sai, "Creating speaker independent ASR system through prosody modification based data augmentation," *Pattern Recognit. Lett.*, vol. 131, pp. 213–218, Mar. 2020, doi: [10.1016/j.patrec.2019.12.019](https://doi.org/10.1016/j.patrec.2019.12.019).
- [45] V. Bhardwaj, V. Kukreja, and A. Singh, "Usage of prosody modification and acoustic adaptation for robust automatic speech recognition (ASR) system," *Revue d'Intell. Artificielle*, vol. 35, no. 3, pp. 235–242, Jun. 2021, doi: [10.18280/ria.350307](https://doi.org/10.18280/ria.350307).
- [46] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition—The 'Ethiopian' system for the SLT 2021 children speech recognition challenge," Nov. 2020, *arXiv:2011.04547*. Accessed: Jun. 15, 2023.
- [47] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral modification based data augmentation for improving end-to-end ASR for children's speech," Mar. 2022, *arXiv:2203.06600*. Accessed: Jun. 12, 2023.
- [48] J. J. Burred, E. Ponsot, L. Goupil, M. Liuni, and J.-J. Aucouturier, "CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition," *PLoS ONE*, vol. 14, no. 4, Apr. 2019, Art. no. e0205943, doi: [10.1371/journal.pone.0205943](https://doi.org/10.1371/journal.pone.0205943).
- [49] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," Nov. 2017, *arXiv:1710.10467*. Accessed: Jan. 16, 2023.
- [50] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011, doi: [10.1109/TASL.2010.2064307](https://doi.org/10.1109/TASL.2010.2064307).
- [51] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4052–4056, doi: [10.1109/ICASSP.2014.6854363](https://doi.org/10.1109/ICASSP.2014.6854363).
- [52] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333, doi: [10.1109/ICASSP.2018.8461375](https://doi.org/10.1109/ICASSP.2018.8461375).
- [53] Zchelly. (Jun. 15, 2020). *Usage of Speaker Embeddings for More Inclusive Speech-to-Text—H2020 COMPRISE*. Accessed: May 5, 2023. [Online]. Available: <https://www.comprish2020.eu/usage-of-speaker-embeddings-for-more-inclusive-speech-recognition/>
- [54] Resemble AI. (Jan. 15, 2023). *Resemble-AI/Resemblyzer*. Accessed: Jan. 16, 2023. [Online]. Available: <https://github.com/resemble-ai/resemble-ai>
- [55] *P800: Methods for Subjective Determination of Transmission Quality*. Accessed: Jan. 17, 2023. [Online]. Available: <https://www.itu.int/rec/T-REC-P800-199608-I>
- [56] R. Jain, A. Barcovschi, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A WAV2VEC2-based experimental study on self-supervised learning methods to improve child speech Recognition," *IEEE Access*, vol. 11, pp. 46938–46948, 2023, doi: [10.1109/ACCESS.2023.3275106](https://doi.org/10.1109/ACCESS.2023.3275106).



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



ANDREI BARCOVSCHI received the B.Eng. degree in electronic and computer engineering from the University of Galway, in 2020, and the M.Sc. degree in artificial intelligence from the National University of Ireland Galway (NUIG), in 2021. He is currently pursuing the Ph.D. degree in artificial intelligence with the University of Galway, researching speech synthesis and conversion technologies, text-to-speech, and speech-to-text. His research interests include machine learning and artificial intelligence topics.



machine learning and artificial intelligence specifically in the domain of speech understanding, text-to-speech, speaker recognition, and automatic speech recognition.

RISHABH JAIN (Graduate Student Member, IEEE) received the B.Tech. degree in computer science and engineering from the Vellore Institute of Technology (VIT), in 2019, and the M.S. degree in data analytics from the University of Galway, Ireland, in 2020, where he is currently pursuing the Ph.D. degree. He is also a Research Assistant with the University of Galway under the Data-center Audio/Visual Intelligence on-Device (DAVID) Project. His research interests include



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

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

Dr. Cucu was awarded the Romanian Academy Prize “Mihail Drăgănescu,” in 2016, for outstanding research contributions in Spoken Language Technology, after developing the first large-vocabulary automatic speech recognition system for the Romanian language.



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

...