IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

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

Digital Object Identifier 10.1109/ACCESS.2023.3275106

RESEARCH ARTICLE

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

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

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

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

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

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

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

I. INTRODUCTION

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

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

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

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

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

A. RELATED WORKS

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

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

Each of these methods show some improvements in child ASR accuracy, however, they still require corresponding labeled annotations to speech. A recent review of child ASRs [21] determined that most of these SOTA methods are supervised learning approaches. The authors in [29] show the performance of various supervised learning approaches for ASR in child speech. They compared the performance of end-to-end ASR systems with that of Deep Neural Network-Hidden Markov Model (DNN-HMM) hybrid systems. Another paper [30] studied the performance of Factored Time Delay Neural Networks (TDNN-F) with traditional and SOTA systems for ASR of child speech. These supervised approaches rely on labeled child speech data during training for the task of ASR.

As there is a distinct lack of labeled child speech data compared to adult, approaches that utilize unsupervised [31] and self-supervised learning [1] were explored for this paper. Therefore, the goal of this work is to present a method to incorporate unlabeled child speech data into the training procedure of a typical ASR model while also making use of abundant, labelled, and unlabeled adult speech data to improve the overall accuracy of ASR models on child speech.

B. SELF-SUPERVISED LEARNING FOR CHILD ASR

Self-supervised learning (SSL) has emerged as a paradigm to learn general data representations from substantial amounts of unlabeled examples allowing one to then fine-tune models on small amounts of labeled data. The use of SSL for child ASR was first seen at Interspeech2021, where a model using SSL [32] received first place for non-native child speech challenge. A similar use case [24] was also presented in the SLT 2021 children speech recognition challenge [33]. Another approach is used in [34], where the author uses a bidirectional unsupervised model pretraining with child speech ASR. After reviewing various approaches to SSL, wav2vec2 [1] was chosen for this paper. Wav2vec2 shows that using SSL for the task of ASR provides improvements over SOTA supervised learning approaches.

At the time of working on this paper, many applications of the wav2vec2 model for child ASR were observed. The authors in [35] propose the use of a transformer model pretrained on adult speech to achieve SOTA results on children's dataset. Reference [36] a comparison between different SSL approaches for child speech recognition tasks. In [37], authors proposed a Domain Responsible Adaptation and Fine-Tuning (DRAFT) framework to address the domain shift between adult speech used for pretraining and child speech used for finetuning. They use wav2vec2 along with other SSL methods to examine the cross-domain transfer between different children's datasets.

This paper explores various pretraining and finetuning configurations with different combinations of adult and child speech datasets using wav2vec2 speech representations. Three child speech datasets were used in this study. These datasets were cleaned and preprocessed to make them usable for ASR. We also report the best results on different child speech validation. The ideal data requirement for pretraining and finetuning in a low-data scenario was also explored in this paper by observing the relation/pattern of performance

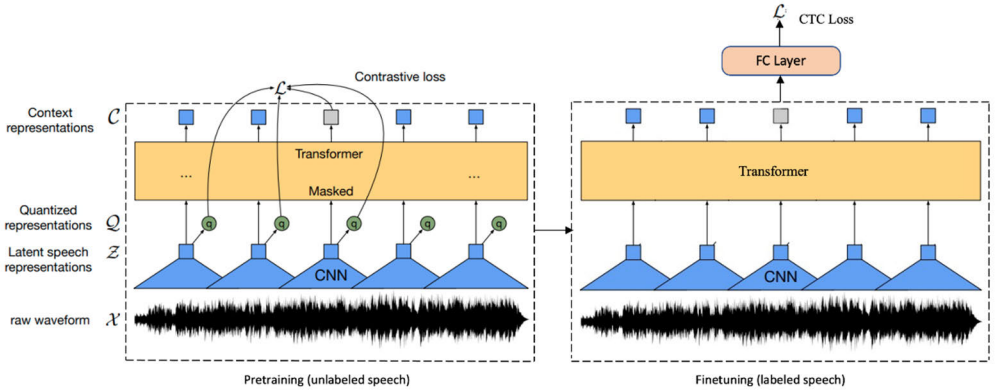


FIGURE 1. Pretraining and finetuning steps in Wav2vec2 (from [1]).

in different datasets used. The rest of this paper is organized as follows: Section II describes the model architecture. Section III introduces the datasets used for this paper. Section IV includes the codebase and experiments. Section V describes the results. Conclusions are presented in Section VI.

II. TRAINING METHODOLOGY FOR SSL

The wav2vec2 model [1] is used to extract speech representations from raw audio files in a self-supervised learning scenario and use these representations for ASR-specific tasks. Wav2vec2 is used in this paper as it can achieve SOTA results when trained on a large amount of unlabeled speech data and finetuned on labeled data as small as 10 minutes. This is ideal for our task, as it is much easier to obtain significant amounts of unlabeled child speech data than gather accurately labeled data.

As it is a two-step training method (See Figure 1), the first step includes a pretraining step in which the model is trained with a large amount of unlabeled data. The second step includes finetuning on labeled data using Connectionist Temporal Classification (CTC) loss [38] for downstream ASR tasks. As the model learns SSL speech representation in pretraining, it can be trained using large quantities of unlabeled speech data and can be finetuned with only a small amount of labeled data. This way, the problem of scarcity of child speech is solved as we can train the 'pretraining' model with a combination of unlabeled speech data and it can also be used to learn speech representations from adult speech datasets making use of the abundant adult speech data.

A. PRETRAINING

The pretraining stage of the wav2vec2 model consists of a feature encoder, context network, and quantization module. The CNN feature extractor takes the raw audio waveform as input and passes it through a series of 1D convolutional layers to extract high-level representations from the waveform. The output of the feature extractor is a sequence of feature vectors that represent the input waveform. The context network is a transformer-based encoder which takes this sequence of feature vectors and processes them using a stack of transformer layers. The transformer layers in wav2vec 2.0 use a self-attention mechanism allowing the model to capture long-range dependencies in the input data. The quantization module consists of a codebook of fixed vectors, where each input feature vector is assigned to the closest codebook vector. Gumbel softmax function [39] is used to choose the quantized representation from multiple codebooks. After quantization, the discrete symbols are passed through a transformer encoder, which learns to encode the sequence of symbols into a fixed-length representation that can be used for downstream tasks such as speech recognition. Since the process involves mapping continuous values to discrete values, it makes the model to be more efficient for training and inference.

The contrastive loss function in Wav2vec2 is applied after the quantization is performed. It is used to train the model to produce embeddings that capture useful features of speech signals. This is followed by a diversity loss which encourages similar feature vectors to be closer together and dissimilar feature vectors to be farther apart. By minimizing these losses, Wav2vec2 can learn to produce embeddings that are effective for downstream speech recognition tasks.

Experiments' configurations are provided as the BASE and LARGE models. The configurations differ in transformer block size but use the same size for the encoder. The feature encoder contains seven blocks with each block having strides of (5,2,2,2,2,2,2) and kernel widths of (10,3,3,3,3,2,2) and output temporal convolution of 512 channels. The context network of the BASE model contains 12 transformer blocks, each block with a 512-dim model, 8 attention heads, and a 2048-dim feed-forward inner layer, while the LARGE model contains 24 transformer blocks with model dimensions 1024, inner dimensions 4096, and 16 attention heads. We use 4 NVIDIA Tesla V100 GPUs to pretrain the model. Model pretraining was optimized using ADAM [40]. During the first 8% of updates, the learning rate warms up to a peak of 5×10^{-4} for BASE and 3×10^{-4} for LARGE, and then it linearly decays. We use both BASE and LARGE models

according to dataset size used for pretraining. BASE models contain 93M parameters and LARGE models contain 317M parameters.

B. FINETUNING

For finetuning, 29 target letters were used (from the Librispeech dataset) as provided by the authors in wav2vec2 [1]. Models are optimized by minimizing CTC loss [38] for ASR task. A modified version of SpecAugment [25] is applied as masking to timestamps and channels to reduce the overfitting and improve the recognition robustness. We fine-tune on one V100 GPU. For the first 1000 updates, only the final output classifier was trained, after which the Transformer block was also trained. The feature encoder was frozen during finetuning training. We also use different finetuning configurations depending on the size of finetuning datasets. The hyperparameters are kept the same as provided by the wav2vec2 authors [1]. The learning rate changes according to the dataset size as documented by the authors of wav2vec2 [1].

As the goal of this study is to evaluate the performance of self-supervised speech representations, it was decided not to incorporate a language model in this research. Additionally, previous research has shown that the best results for children's ASR systems were achieved without the use of an external language model [29]. Language model adaptation for child speech is also an unexplored research area. Child speech would require a specialized trained language model for best results. As there isn't any definitive publicly available language model for child speech, we consider this as a part of the future research topic.

III. DATASET DESCRIPTION AND USAGE

The datasets are divided according to their usage. The child speech data used in this paper include MyST Corpus [41], CMU_Kids [42] and PF-STAR [43]. Adult Speech datasets include Librilight [44], LibriTTS [45], and LibriSpeech [46].

A. DATASET DESCRIPTION

Below we provide a description of the datasets used in this paper:

1) LIBRISPEECH [46]

Librispeech is an adult speech dataset with approximately 1000 hours of recorded audio with a sampling rate of 16Khz. The data is derived from read audiobooks from the LibriVox project. The data is carefully segmented, aligned, and used popularly in speech research.

2) LIBRILIGHT [44]

Librilight is an adult speech dataset used as a benchmark for training speech recognition systems with limited or no supervision. It contains 60,000 hours of unlabeled adult speech extracted from audiobooks. It was mentioned in the wav2vec2 paper [1] and used by the authors.

3) LibriTTS [45]

The LibriTTS dataset is a large-scale dataset for training TTS models and is a subset of the Librispeech dataset. It consists of approximately 560 hours of high-quality audio and text transcriptions from audiobooks. This dataset is used here for inference over adult speech as it is a clean and noise-free dataset. The 'dev-clean' segment of the LibriTTS dataset which contains over 8.9 hours of clean adult speech. It is also widely used as a baseline in the validation of ASR and TTS experiments.

4) MY SCIENCE TUTOR (MySt) CHILD SPEECH [41]

The MyST (My Science Tutor) Children's Speech Corpus consists of 393 hours of American English children's speech with a total of 228,874 utterances. The speech was collected from 1371 third, fourth and fifth-grade students. 45% of the utterances have been transcribed at the word level amounting to 197 hours. This dataset is used in this paper as it's the largest open-source corpus of child speech available for research use.

5) PF-STAR CORPUS OF BRITISH ENGLISH CHILD SPEECH [43]

This corpus contains British English child speech from 158 children aged 4 to 14 years. The recordings are divided into a training set (7.5 hours), an evaluation set (1 hour) and a test set (5.6 hours). The corpus was collected at three locations: a university laboratory and two primary schools. It contains both read and spontaneous child speech with transcriptions.

6) CMU KIDS [42]

CMU KIDS Corpus contains read-aloud sentences by children. It was created to provide training data for the SPHINX II automatic speech recognizer at Carnegie Mellon University. It contains 9 hours of American English child speech. The dataset contains 24 male and 52 female speakers having a total of 5180 utterances.

B. DATASET CLEANING AND PROCESSING

All speech data was converted into a 16-bit mono channel with a 16Khz sampling rate, wherever required. All the transcriptions were cleaned and normalized to remove abbreviations, punctuations, whitespaces, etc. and all the characters were changed to uppercase. All the non-linguistic annotation symbols (in child speech datasets) such as "<unk>, sil, hmm, <breath>, <noise>, <indiscernible>, [ze-], [cham-], [***ision], etc." were removed and only alphanumeric characters were retained in the transcript. This was done for all the labeled data used in this paper. Child datasets required further cleaning and pre-processing as follows:

1) MYST CLEANUP

We use the transcribed portion of MyST dataset containing over 197 hours of speech data presented in .trn file

format. The MyST dataset contained a lot of noisy and non-meaningful sentences such as:

- <silence> I'm i don't know <noise> actually
- <whisper> sending go back (*)
- <whisper> what's this one <side_speech> it's an
- give me that <indiscernible> a circuit is a pathway
- <laugh> yeah yeah

The content between '<' and '>' tags were removed from all the transcriptions along with the tags themselves. All the cleaned text files were saved in a .txt format. On further inspection, it was observed that samples below 10 seconds in length generally contained non-meaningful, noisy speech, and data above 20 seconds would lead to GPU running out of memory. Therefore, 10-20 seconds long speech samples from transcribed MyST were selected for finetuning. A final cleaning was performed by manually removing some of the non-meaningful utterances by listening to audio files and going through the transcripts, which amounted to a total of 65 hours of clean data. The data was then randomly split into two groups having 55 hours of data for training and 10 hours for testing as can be seen in Table 1.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="(unknown)" audio_filename="digits1" version="2" version_date="031105">
<Episode>
<Section type="report" startTime="0" endTime="36.497">
<Turn startTime="0" endTime="36.497">
<Sync time="0"/>
sil
<Sync time="0.985"/>
five
<Sync time="1.735"/>
two
<Sync time="2.289"/>
four
<Sync time="2.852"/>
sp
<Sync time="3.445"/>
seven
<Sync time="4.098"/>
sp
<Sync time="4.258"/>
five
<Sync time="4.727"/>
nine
<Sync time="5.289"/>
sp
<Sync time="5.883"/>
one
<Sync time="6.414"/>
oh
<Event desc="error_wrong_word(s)" type="noise" extent="instantaneous"/>
<Sync time="6.914"/>
one
<Sync time="7.289"/>
sil
```

FIGURE 2. Example of '.trs' file in the pfstar dataset. The content in this image was segmented into 'five two four', 'seven', 'five nine', and 'one oh one'. The image is provided to show an example of how transcripts data were made available using '.trs' transcriber old format.

2) PFSTAR CLEANUP

The PFSTAR corpus also contained a lot of non-meaningful utterances and noisy data samples. The dataset comes with '.trs' transcription files, containing time-aligned text information (see Figure 2). These timestamps were used to further segment the data into small audio chunks and remove noise from the dataset. The 'sp' tag from the transcription was used to divide the long transcripts into smaller segments. The corresponding time information was used to segment the long audio files into smaller chunks using FFmpeg[1] and Python. The audio files from the PFSTAR dataset which were 30-70 seconds long were segmented into smaller audio chunks of 5-20 seconds in duration. This segmentation led to 12 hours of clean, usable PFSTAR data, which was further divided into 2 sets: PFS_10h with 10 hours of data (for training) and PFS_test with 2 hours of data (for inference). The final audio data was saved in .wav format and transcriptions in .txt format.

[1]FFmpeg: https://ffmpeg.org/

3) CMUKIDS CLEANUP

CMU_Kids dataset also contains a lot of noisy and incomprehensible child speech. The transcriptions are provided in a '.trn' file format and audio files in a '.sph' format. The data was cleaned in a similar way to MyST by removing all the unrequired tags and non-textual information from the transcripts. For example, ''they [begin_noise] kept a few [end_noise] butterflies in [noise]'' was converted to ''they kept a few butterflies in''. A few more examples can be seen below:

- [begin_noise] cages [end_noise] to lay more eggs [noise] [sil]
  -> cages to lay more eggs
- a [begin_noise] blue butterfly [end_noise] /F L R UW/ [human_noise] flew by [human_noise] [human_noise]
  -> a blue butterfly flew by

The cleaned dataset contained all the audio files in '.wav' format and all transcribed speech in '.txt' format as needed for our training. The total amount of CMU_Kids dataset amounted to 9 hours which was used during inference only.

C. DATASET USAGE

The dataset usage is mentioned in Table 1. The 'Usage' column indicates whether the dataset was used for

TABLE 1. Dataset description for pretraining, finetuning and inference.

Usage	Dataset	Duration	Type
Pretraining [Unlabeled data]	MyST_complete	393 hrs	Child
	Librispeech	960 hrs	Adult
	Libri-light	60k hrs	Adult
Finetuning [Labeled data]	MyST_10m	10 mins	Child
	MyST_1h	1 hr	Child
	MyST_10h	10 hrs	Child
	MyST_55h	55 hrs	Child
	PFS_10m	10 mins	Child
	PFS_1h	1 hr	Child
	PFS_10h	10 hrs	Child
	LS_10m	10 mins	Adult
	LS_100h	100 hrs	Adult
	LS_960h	960 hrs	Adult
Inference [Labeled data]	MyST_test	10 hrs	Child
	PFS_test	2 hrs	Child
	CMU_Kids	9 hrs	Child
	LibriTTS 'dev-clean'	8.9 hrs	Adult

TABLE 2. Group-A: WER for different pretraining (Adult speech datasets) and finetuning (Adult speech dataset) experiments on the MYST, PF-STAR, CMU KIDS and LIBRITTS 'dev-clean' datasets.

Group	Model ID	Pretraining Model Configuration	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMU_KIDS	WER dev_clean
GROUP - A	1	BASE	Librispeech	LS_10m	31.48	30.05	33.38	15.90
	2			LS_100h	17.82	15.96	18.73	4.16
	3			LS_960h	**15.41**	**11.20**	**16.33**	**3.40**
	-	*Average (Group – A, BASE)*			*21.57*	*19.07*	*22.81*	*7.82*
	4	LARGE	Librilight	LS_10m	26.47	27.14	29.37	15.35
	5			LS_100h	13.15	11.63	16.18	3.79
	6			LS_960h	**12.50**	**8.56**	**14.85**	**3.28**
	-	*Average (Group – A, LARGE)*			17.37	15.78	20.13	7.47

pretraining, finetuning, or inference. The 'Type' column specifies whether the dataset consists of child or adult speech. Dataset name is mentioned in 'Dataset' column while amount (in hours/minutes) is mentioned under 'Duration' column.

Pretraining datasets only consists of audio files and doesn't require any transcript/labelled data during training. Finetuning data consists of audio files along with labelled transcripts. The size of the finetuning datasets was chosen as instructed in wav2vec2 [1], and to keep it consistent with their methodology. A similar distribution was maintained for finetuning with child speech datasets (wherever possible). The data was segmented randomly for creating various finetuning subsets.

IV. CODEBASE AND EXPERIMENTS

A. CODEBASE AND HYPERPARAMETERS

The wav2vec2 implementation provided by the fairseq[2] framework is used for our experiments. Hyperparameters were kept the same for both BASE and LARGE pretraining configurations as provided by the wav2vec2 authors. Finetuning configurations were also kept consistent with the finetuning dataset size used. Data cleaning and data processing scripts were created using FFmpeg and Python-based tools such as pydub and scipy. All the training checkpoints are made available on our GitHub page[3] and can be used directly with the model implementation from fairseq. See note[4] for more information on data cleaning scripts and dataset availability.

B. EXPERIMENTS

Experiments were divided into five groups, Group-A, B, C, D and E. ASR performance is measured in terms of Word Error Rate (WER) on different adult and child speech datasets. Child speech datasets used in inference include unseen MyST_test, PFS_test and CMU_Kids, and adult speech dataset include LibriTTS 'dev-clean'. These datasets are common for all groups during inference tests. All the groups of experiments (except Group-C) use two model configurations, namely BASE and LARGE. The BASE configuration includes 960 hours of Librispeech pretraining data and the LARGE configuration includes 60k hours of Librilight data, which is 60 times as much pretraining data as in the BASE configuration. This enables an assessment of the importance of the original training data size for the wav2vec2 model.

For Group-A (Table 2), the finetuned checkpoints provided by the wav2vec2 repository were used for inference. Each of the BASE and LARGE configurations were finetuned with 10 minutes, 100 hours, and 960 hours of Librispeech. For Group-B (Table 3), the pretrained model is finetuned with 10 minutes, 1 hour, 10 hours, and 55 hours of MyST child speech data. In Group-C (Table 3), the Librispeech and MyST datasets having 960 hours of adult speech and 393 hours of child speech data, respectively, are used for pretraining. The model is then finetuned over different amounts of the MyST dataset (similar to Group-B). We only use BASE configuration for this experiment. Group-D (Table 4) uses PFSTAR dataset for finetuning instead of the MyST dataset, and both BASE and LARGE configuration are finetuned with 10 minutes, 1 hour and 10 hours of PFSTAR child speech dataset. Group-E (Table 5) uses a mix of different datasets in the finetuning. A mix of the MyST_55h, PFS_10h, and LS_960 datasets was used. Finetuning mix included LS_960h+MyST_55h, LS_960h+PFS_10h, MyST_55h+PFS_10h and LS_960h+MyST_55h+ PFS_10h. These experiments were performed to see the cross-domain correlation in WER across different finetuning datasets.

Note that we did not train any models from scratch with child speech data alone as there is not sufficient publicly available child speech data to learn any meaningful speech representations from child speech alone. This is discussed in more detail in section V.

V. RESULTS AND DISCUSSION

A. MAIN RESULTS FROM THE GROUP EXPERIMENTS

Results from group experiments are presented in Tables – 2, 3, 4, and 5, with lowest WERs highlighted in bold.

[2]https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

[3]https://github.com/C3Imaging/childASR_w2v2

[4]**Note:** We only make the basic data cleaning scripts available in the GitHub. Researchers trying to replicate our work can email us and get access to other research material. For access to respectively cleaner versions of datasets used in this paper, researchers can buy their own license for the original datasets (where required), and on providing proof of that license, can get access to our 'clean' versions.

TABLE 3. Group-B and Group-C: WER for different pretraining (adult and child speech datasets) and finetuning (MyST child speech dataset) combinations on the MYST, PF-STAR, CMU KIDS and LIBRITTS 'dev-clean' datasets.

Group	Model ID	Pretraining Model Configuration	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMU_KIDS	WER dev_clean
GROUP-B	7	BASE	Librispeech	MyST_10m	28.84	41.34	34.18	21.45
	8			MyST_1h	18.75	31.84	23.13	13.91
	9			MyST_10h	13.46	28.68	19.59	10.94
	10			MyST_55h	**8.13**	**14.77**	**16.47**	**7.72**
	-	*Average (Group – B, BASE)*			*17.29*	*29.16*	*23.34*	*13.51*
	11	LARGE	Librilight	MyST_10m	33.01	44.36	39.91	46.45
	12			MyST_1h	14.91	26.21	18.74	11.59
	13			MyST_10h	12.92	25.05	17.72	10.04
	14			MyST_55h	**7.51**	**12.46**	**15.25**	**6.43**
	-	*Average (Group – B, LARGE)*			*17.08*	*27.02*	*22.91*	*18.62*
GROUP-C	15	BASE	Librispeech MyST_Complete	MyST_10m	29.16	45.71	37.56	35.39
	16			MyST_1h	21.89	38.53	29.03	20.45
	17			MyST_10h	16.18	32.95	25.06	16.83
	18			MyST_55h	**10.34**	**25.47**	**23.15**	**13.48**
	-	*Average (Group – C, BASE)*			*19.39*	*35.67*	*28.7*	*21.53*

TABLE 4. Group-D: WER for different pretraining (adult speech datasets) and finetuning (PFstar child speech dataset) combinations on the MYST, PF-STAR, CMU KIDS and LIBRITTS 'dev-clean' datasets.

Group	Model ID	Pretraining Model Configuration	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMU_KIDS	WER dev_clean
GROUP-D	19	BASE	Librispeech	PFS_10m	35.91	16.43	33.53	30.43
	20			PFS_1h	33.52	7.36	29.55	16.61
	21			PFS_10h	**31.86**	**3.48**	**27.49**	**13.95**
	-	*Average (Group – D, BASE)*			*33.76*	*9.09*	*30.19*	*20.33*
	22	LARGE	Librilight	PFS_10m	37.10	16.78	35.13	23.85
	23			PFS_1h	30.81	14.19	28.54	21.89
	24			PFS_10h	**27.17**	**3.50**	**21.35**	**11.60**
	-	*Average (Group – D, LARGE)*			31.69	11.49	28.34	19.11

TABLE 5. Group-E: WER for different pretraining (adult datasets) and finetuning (adult and child speech datasets) combinations on the MYST, PF-STAR, CMU KIDS and LIBRITTS 'dev-clean' datasets.

Group	Model ID	Pretraining Model Configuration	Pretraining dataset	Finetuning dataset	WER MyST_test	WER PFS_test	WER CMUKIDS	WER dev_clean
GROUP-E	25	BASE	Librispeech	LS_960h, MyST_55h	8.18	12.17	14.12	1.24
	26			LS_960h, PFS_10h	15.42	3.74	15.31	1.41
	27			MyST_55h, PFS_10h	**7.94**	**2.91**	15.97	7.64
	28			LS_960h, MyST_55h, PFS_10h	8.13	3.12	**13.76**	**1.20**
	-	*Average (Group – E, BASE)*			*9.91*	*5.48*	*14.79*	*2.87*
	29	LARGE	Librilight	LS_960h, MyST_55h	8.06	9.31	13.20	1.34
	30			LS_960h, PFS_10h	13.18	3.17	13.19	**1.32**
	31			MyST_55h, PFS_10h	**7.42**	**2.99**	14.18	5.79
	32			LS_960h, MyST_55h, PFS_10h	8.17	3.33	**12.77**	1.40
	-	*Average (Group – E, LARGE)*			*9.2*	*4.7*	*13.33*	*2.4*

1) GROUP-A (TABLE-2)

In this group, adult datasets are used in both pretraining and finetuning. All models show a pattern of decreasing WER with an increase in the size of the finetuning dataset. It can also be observed that there is not a large difference in WER between BASE and LARGE models even though the LARGE model uses 60 times more training data.

2) GROUP-B (TABLE-3)

All the models in Group-B, finetuned with different amounts of MyST data, attained lower WERs on the child speech in comparison with Group-A experiments. A similar trend of decreasing WER can be observed with an increase in finetuning data.

3) GROUP-C (TABLE-3)

Group-C experiments were designed similar to Group-B (see Table-3). The objective was to investigate whether adding child speech dataset in the pretraining have any impact on the model performance. Comparing to the BASE models from Group-B, the WERs on all test sets increased in Group-C. Therefore, using child speech in pretraining was not considered for Group-D and Group-E experiments.

4) GROUP-D (TABLE-4)

In this group, the PFSTAR dataset was used for fine-tuning. The model's performance also improves as the size of the finetuning dataset increases. The best results, as might be expected, are on PFS_test while results on the other test datasets are less impressive.

5) GROUP-E (TABLE-5)

Group-E used LS_960h, PFS_10h and MyST_55h in various finetuning combinations as these datasets gave the best WER in previous finetuning experiments. Group-E models outperformed all the previous models and gave the best WER for all the inference datasets.

B. DISCUSSION OF RESULTS

Group-A (Table 2) results provide a baseline where only adult speech data is used for pretraining and finetuning. The relative improvements due to finetuning with adult speech are similar across all of the child test datasets, indicating that large adult speech datasets provide similar levels of improvement on different child speech validation. We can draw three additional conclusions. Firstly, there is less than a 3% variation in WER between BASE and LARGE wav2vec2 models across all the test datasets, so the LARGE model is only useful where optimal performance is needed, and BASE models are ideal for low resources scenario. Secondly, the improvement between finetuning with 10 minutes of adult speech data and 100 hours is much more significant than the improvement between 100 hours and 960 hours. There is only a 3% average WER difference between LS_100h and LS_960h finetuning, suggesting 100 hours of adult speech is ideal for finetuning.

Next, after introducing various amounts of child speech data for fine-tuning in Group-B (Table 3), it is noted that smaller amounts of child speech data result in better improvements in WER. It is clear that as little as 1 hour of child speech can have similar improvements to 100 hours of adult speech. Similarly, 10 hours of child speech shows similar improvements as 960 hours of adult speech. However, we also note a significant domain mismatch across the test datasets as the improvements on PFS_test and CMU_Kids are significantly weaker than for MyST_test. An overarching conclusion here might be that 1 hour of child speech is equivalent to 100 hours of adult speech where there is strong domain alignment between the finetuning and test speech. Lastly, using LARGE model for finetuning with only a small amount of child speech (e.g., 10 mins) may be detrimental due to domain mismatch between pretraining and finetuning datasets. Again, there is a relatively small performance improvement between BASE and LARGE models.

The Group-C (Table 3) experiments add the MyST_Complete dataset to the pretraining. Performance is poorer than with adult speech only, highlighting the limitations of pretraining data with the noisy and non-linguistic child speech in the MyST_Complete corpus. Further investigation is needed to understand this impact of child speech data on the pretraining; however, it will require a much cleaner and larger child speech dataset.

Group-D (Table 4) experiments are equivalent to Group-B (Table 3) but use the PFSTAR dataset for fine-tuning. As this dataset is smaller than MyST, only 10 minutes, 1 hour and 10 hours of speech can be used for fine-tuning. The key takeaway here is that PFS_test results improve even more significantly than MyST_test in Group-B, but the other child speech test datasets barely show any improvement. Clearly there is a significant domain mismatch between PFSTAR dataset with British English dialect and the two other child-speech datasets with American English dialect. PFSTAR was also recorded in a much cleaner environment. This shows that properties like dialect, accent and acoustic characteristics can impact the performance of the ASR model. Interestingly, MyST and PFSTAR finetuning (from Group B and D) shows similar WER on LibriTTS dev-clean implying that child speech datasets with distinct properties perform similarly when used for adult speech recognition.

Finally, for the Group-E experiments (Table 5), where a mix of adult and child datasets are used, we find that finetuning on the two child speech datasets, MyST_55h and PFS_10h gives the best results with WER rates of 7.91 and 2.94 on the respective tests datasets, MyST_test and PFS_test. Performance for CMU_Kids is significantly weaker at 15.97. Clearly, when the finetuning data has a good domain match with the tests data then SOTA WER rates can be achieved through finetuning with approximately 65 hours of child speech data. The BASE and LARGE configurations in Group-E show an absolute difference of 0.84 WER suggesting that performance is similar for both configurations when cross-domain datasets combinations are used in finetuning.

Interestingly, using smaller amounts of child speech can provide significant improvements in WER accuracy as compared with large amount of adult speech. This study provides a baseline for future studies. While the results of this study provide a comprehensive analysis of different

TABLE 6. Previous SOTA results on the MYST, PF-STAR, and CMU_KIDS datasets.

SOTA Papers	Method Type	Training Data (hrs)	Inference data (hrs)	WER MyST	WER PFSTAR	WER CMU_Kids
TDNN-F + Augmentation [30]	Supervised	6.34	2.76	-	-	16.01
Hybrid HMM-DNN Transfer Learning [28]	Supervised	6.26	2.45	-	-	19.33
DRAFT [37]: ○ WAV2VEC2 ○ HuBERT	Self-Supervised	197	13	 16.70 16.53	-	-
Transformer + CTC + Greedy [29]	Supervised	197	13	16.01	-	-
W2V2 + source-filter warping + LM [35]	Self-Supervised	11.2	2.5		4.86	

finetuning techniques for child ASR, additional conclusions can be drawn by comparing different experiments.

C. THIS WORK IN THE CONTEXT OF PREVIOUS CHILD SPEECH ASR APPROACHES

As commented in the Introduction, the publicly available child speech datasets are small in comparison to well-established adult speech datasets and audio quality is poor in comparison. Further, if the full datasets are used to build randomized test datasets, then many of the data samples will be of very variable quality. Thus, previous authors have adopted various approaches to clean and utilize the data but due to lack of standardized approach, it would not be fair to make any direct comparisons.

Our best results using the SSL approach show potential for significant improvement over the previously reported results on the same dataset as shown in Table 6. Our trained models achieved the best WER of **7.42** on the MyST_test dataset, **2.91** on the PFSTAR, PFS_test dataset (reaching human level performance) and **12.77** on the CMU_Kids dataset, as compared to the previously reported results from [28], [29], [30], [35], and [37]. Our detailed explanations of how the test datasets were 'cleaned' for this work should further provide researchers with a useful basis for future comparisons.

VI. CONCLUSION

In this work, the wav2vec2 self-supervised training approach is adapted with different mixes of pretraining and finetuning datasets to provide a methodology to improve the accuracy of child speech recognition. A combination of adult and child speech datasets is used to determine the data requirements for improving child speech recognition. Experiments were designed to evaluate the relative performance on the in-domain MyST and PFSTAR datasets, the out-of-domain CMUKIDS dataset while using the LibriTTS dev-clean dataset as a reference adult speech dataset. The best results were obtained where the model was pretrained on adult data and fine-tuned on a combinations of child speech datasets. The best WER rates (7.42 on MyST_test, 2.91 on PFS_test, 12.77 on CMU_Kids) are comparable with the best SOTA results available currently in the literature.

A model pretrained with adult speech data can best learn the speech features as compared to a model including both adult and child speech in pretraining. In particular, adding a low-quality dataset such as the MyST child speech dataset in pretraining reduced the performance of the ASR model across all test datasets. Significant domain variations were also evident between the MyST, CMU_Kids and PFSTAR datasets with the latter being of notably better quality. Qualitatively we can say that MyST and CMU_Kids are more closely aligned than the PF-Star dataset. When a cross-domain mix of child speech is used for fine-tuning (e.g., model 27 or model 31) then the optimal results are achieved. For a model finetuned with single or multiple child/adult speech data, WER increases over the dataset with similar distribution as finetuning dataset.

The BASE configuration of wav2vec2, which is pretrained with 60 times less data than the LARGE configuration is effective for a low-data scenario. In fact, the improvements achieved through using the LARGE configuration were typically only a few percent and hardly seem to justify the large increase in computational resources needed to train. As for finetuning, we can say that 100 hours of adult speech finetuning data offer a practical trade-off between computational effort and ASR accuracy. Finetuning with as little as 10 hours of child speech data provided better improvement over models finetuned with 960 hours of adult speech. Optimal results are achieved using in the order of 65 hours of cross-domain child speech (a mix of MyST and PFSTAR).

For future work, these models can be used to transcribe additional child speech data from the unlabeled MyST dataset and a range of additional unlabeled datasets. It would also be interesting to investigate the potential of generative data augmentation models [47] to provide additional synthetic child speech samples and a wider variety of child speech for pretraining and finetuning experiments.

ACKNOWLEDGMENT

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

REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, and J. Chen, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[3] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*.

[4] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6124–6128.

[5] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[6] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "ContextNet: Improving convolutional neural networks for automatic speech recognition with global context," 2020, *arXiv:2005.03191*.

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[8] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999.

[9] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 1997, pp. 1–4.

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

[11] V. Bhardwaj, M. T. B. Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (ASR) systems for children: A systematic literature review," *Appl. Sci.*, vol. 12, no. 9, p. 4419, Apr. 2022, doi: 10.3390/APP12094419.

[12] R. Peinl and J. Wirth, "Quality assurance for speech synthesis with ASR," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2022, pp. 739–751.

[13] A. Baby, S. Vinnaitherthan, N. Adiga, P. Jawale, S. Badam, S. Adavanne, and S. Konjeti, "An ASR guided speech intelligibility measure for TTS model selection," Jun. 2020, *arXiv:2006.01463*.

[14] V. Kadyan, H. Kathania, P. Govil, and M. Kurimo, "Synthesis speech based data augmentation for low resource children ASR," in *Speech and Computer* (Lecture Notes in Computer Science), vol. 12997. Cham, Switzerland: Springer, 2021, pp. 317–326, doi: 10.1007/978-3-030-87802-3_29.

[15] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 976–989, 2020, doi: 10.1109/TASLP.2020.2977776.

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

[17] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian, "Towards data selection on TTS data for children's speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6888–6892, doi: 10.1109/ICASSP39728.2021.9413930.

[18] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario," in *Proc. Interspeech*, 2020, pp. 4382–4386, doi: 10.21437/Interspeech.2020-1112.

[19] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, "Data augmentation using CycleGAN for end-to-end children ASR," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 511–515, doi: 10.23919/EUSIPCO54536.2021.9616228.

[20] N. Jia, C. Zheng, and W. Sun, "Speech synthesis of children's reading based on CycleGAN model," *J. Phys., Conf.*, vol. 1607, no. 1, Aug. 2020, Art. no. 012046, doi: 10.1088/1742-6596/1607/1/012046.

[21] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2014, pp. 135–140.

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

[23] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *Proc. Interspeech*, Sep. 2016, pp. 1598–1602, doi: 10.21437/INTERSPEECH.2016-1348.

[24] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition—The 'ethiopian' system for the SLT 2021 children speech recognition challenge," Nov. 2020, *arXiv:2011.04547*.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.

[26] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral modification based data augmentation for improving end-to-end ASR for children's speech," 2022, *arXiv:2203.06600*.

[27] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Comput. Speech Lang.*, vol. 63, Sep. 2020, Art. no. 101077, doi: 10.1016/J.CSL.2020.101077.

[28] T. Rolland, A. Abad, C. Cucchiarini, and H. Strik, "Multilingual transfer learning for children automatic speech recognition," in *Proc. 13th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 7314–7320.

[29] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101289.

[30] F. F. Wu, L. P. Garcia, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network," in *Proc. Interspeech*, 2019, pp. 1–5, doi: 10.21437/Interspeech.2019-2980.

[31] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 27826–27839.

[32] G. Xu, S. Yang, L. Ma, C. Li, and Z. Wu, "The TAL system for the INTERSPEECH2021 shared task on automatic speech recognition for non-native children's speech," in *Proc. Interspeech*, 2021, pp. 1294–1298, doi: 10.21437/Interspeech.2021-1104.

[33] F. Yu, Z. Yao, X. Wang, K. An, L. Xie, Z. Ou, B. Liu, X. Li, and G. Miao, "The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 1117–1123.

[34] R. Fan, A. Afshan, and A. Alwan, "Bi-APC: Bidirectional autoregressive predictive coding for unsupervised pre-training and its application to children's ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7023–7027.

[35] J. Thienpondt and K. Demuynck, "Transfer learning for robust low-resource children's speech ASR with transformers and source-filter warping," 2022, *arXiv:2206.09396*.

[36] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards better domain adaptation for self-supervised models: A case study of child ASR," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1242–1252, Oct. 2022, doi: 10.1109/JSTSP.2022.3200910.

[37] R. Fan and A. Alwan, "DRAFT: A novel framework to reduce domain shifting in self-supervised learning and its application to children's ASR," in *Proc. Interspeech*, 2022, pp. 1–5, doi: 10.21437/Interspeech.2022-11128.

[38] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[39] E. Jang, G. Brain, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. ICLR*, 2017, pp. 1–13.

[40] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

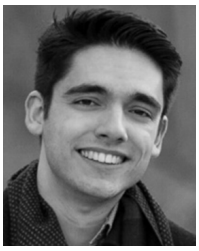
[41] W. Ward, R. Cole, and S. Pradhan, "My science tutor and the myst corpus," Boulder Learn. Inc., 2019.

[42] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids corpus LDC97S63," Tech. Rep. LDC97S63, 1997. [Online]. Available: https://catalog.ldc.upenn.edu
[43] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," 2005.
[44] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for ASR with limited or no supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7669–7673.
[45] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech*, 2019, pp. 1–7, doi: 10.21437/interspeech.2019-2441.
[46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.
[47] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis," *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.

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

ANDREI BARCOVSCHI received the B.Eng. degree in electronic and computer engineering from the University of Galway (prior to 2023: National University of Ireland Galway (NUIG)), in 2020 and the M.Sc. degree in artificial intelligence from NUIG, in 2021. He is currently pursuing the Ph.D. degree in artificial intelligence with the University of Galway. His research interests include speech synthesis and conversion technologies, text-to-speech, and speech-to-text. He is interested in a broad range of machine learning and artificial intelligence topics.

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



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

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

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

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

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