Transfer Learning for Robust Low-Resource Children's Speech ASR with Transformers and Source-Filter Warping

Jenthe Thienpondt, Kris Demuynck

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium

jenthe.thienpondt@ugent.be

Abstract

Automatic Speech Recognition (ASR) systems are known to exhibit difficulties when transcribing children's speech. This can mainly be attributed to the absence of large children's speech corpora to train robust ASR models and the resulting domain mismatch when decoding children's speech with systems trained on adult data. In this paper, we propose multiple enhancements to alleviate these issues. First, we propose a data augmentation technique based on the source-filter model of speech to close the domain gap between adult and children's speech. This enables us to leverage the data availability of adult speech corpora by making these samples perceptually similar to children's speech. Second, using this augmentation strategy, we apply transfer learning on a Transformer model pre-trained on adult data. This model follows the recently introduced XLS-R architecture, a wav2vec 2.0 model pre-trained on several cross-lingual adult speech corpora to learn general and robust acoustic frame-level representations. Adopting this model for the ASR task using adult data augmented with the proposed source-filter warping strategy and a limited amount of in-domain children's speech significantly outperforms previous state-of-the-art results on the PF-STAR British English Children's Speech corpus with a 4.86% WER on the official test set.

Index Terms: speech recognition, children's speech, data augmentation, Transformers

1. Introduction

Automatic speech recognition (ASR) performance on adult speech data has recently improved noticeably due to the availability of large transcribed speech corpora [1, 2] and the development of end-to-end attention-based acoustic models to leverage the available data [3, 4]. However, in low-resource settings, such as children's speech recognition, the performance benefits of these end-to-end models are limited due to the lack of substantial in-domain transcribed data with more traditional approaches such as DNN-HMM ASR models still being competitive [5, 6].

The recent advances of transfer learning in the field of ASR shows promising results on similar low-resource speech recognition tasks [7, 8, 9]. Fine-tuned acoustic frame-level representations from self-supervised models pre-trained with a masking objective on unlabelled adult data can be successfully used for downstream speech recognition applications with a small amount of data [10]. Encouraging results using pre-trained end-to-end models have recently been established for children's speech recognition with limited amounts of transcribed in-domain data [11, 12]. However, these models are pre-trained using large amounts of unlabelled in-domain children's speech, which imposes an important limitation on the usage of this approach. To alleviate this assumption, we attempt to leverage unlabelled adult speech in combination with a data augmentation strategy to tackle the acoustic mismatch between adult and children's speech.

Vocal tract length pertubation (VTLP) [13] is the most established method to close the domain gap between adult and child speakers [14]. VTLP applies a linear warping along the frequency axis of the adult speech spectrum to make it perceptually more similar to children's speech. Other work has gained improvements by generating additional input samples by transforming adult speech into children's speech using a voice conversion model based on cycle GANs [15, 16]. However, voice conversion models require supplementary child data and introduces an extra training stage which needs to be optimized for the downstream task.

In this work, we propose an augmentation technique based on the source-filter model of speech [17]. We argue that the characteristics of the source and filter component of the speech spectrum behave independently in relation to the adult and children's speech domain mismatch. Subsequently, we introduce a data augmentation strategy which applies a warping function with separate configurations for the source and filter component of the input signal. This enables us to use available adult speech to train more robust acoustic models for transcribing children's speech. We develop an end-to-end acoustic model based on the recently introduced XLS-R model [10]. This architecture based on wav2vec 2.0 [8] is pre-trained self-supervised on large cross-lingual corpora of adult speech with the task of predicting quantized units of masked latent speech representations. Using the proposed source-filter warping strategy enables us to leverage available transcribed adult data and fine-tune the model robustly on the children's speech recognition task.

2. Data augmentation

The disparities between adult and children's acoustic characteristics poses some inherit challenges to speech recognition systems for children's speech. Mainly, formants in children's speech are located at higher frequencies and are prone to higher inter-speaker variability due to the shorter and developing vocal tract of children when compared to adults [18]. A shorter vocal tract and subsequent higher fundamental frequency (f_0) also results in undersampling of the spectral envelope due to the widely spaced harmonics [19, 20]. This makes speech recognition methods relying on spectral representations of the input signal less robust when handling children's speech, especially when trained on adult data, which does not exhibit this problem due to a lower average f_0. Other distinctions include age-dependent cognitive abilities leading to more frequent disfluencies and mispronunciations [21]. Several techniques have been proposed to make the spectral representation of adult data more similar to children's speech.

2.1. Vocal tract length pertubation

The most used data augmentation strategy to mimic the spectral characteristics of children is applying VTLP [13] on adult data [14]. This method applies a linear warping function with a random factor η along a range of frequencies covering the significant formants in the spectrum of the signal. In the case of children's speech recognition, η is usually constricted to $\eta > 1$ since the average fundamental frequency and formant locations of children's speech is higher in comparison to male and female adult speech.

2.2. Proposed source-filter warping

In the source-filter model of speech production, speech $y(t)$ is regarded as the convolution of an input signal $s(t)$ and an impulse response $v(t)$, often referred to as the source and filter component, respectively [17]. The resulting equation for speech production in the source-filter model is $y(t) = s(t) * v(t)$. Transferring to the spectral domain, the resulting equation becomes:

$$Y(\omega) = S(\omega)V(\omega) \tag{1}$$

with the convolution turned into multiplication. $S(\omega)$ and $V(\omega)$ indicate the source and vocal-tract filter spectrum, respectively. VTLP applies a warping function along the frequency dimension contained in $Y(\omega)$, resulting in the usage of the same warping coefficient for both the source and filter component. However, we argue that the optimal warping factor to transform the adult spectrum to a child-like spectral representation is distinct for the source and filter element. Therefore, we propose source-filter warping (SFW), a data augmentation strategy which applies a warping function with separate warping coefficients α and β for the source and filter component, respectively.

2.2.1. Spectral envelope estimation

We use an iterative smoothing algorithm along the frequency dimension of the power spectrum to estimate the spectral envelope. This reduces the computational complexity as opposed to methods such as cepstral windowing [22] and linear predictive coding (LPC) [23]. Given that Y_i represents the power spectrum at frequency location i after applying the short-time Fourier transform (STFT) on the input waveform, the corresponding spectral envelope V_i is estimated iteratively with:

$$V_i = max(Y_i, V_{i-1} + \gamma(Y_i - V_{i-1})) \tag{2}$$

with $V_0 = Y_0$ and γ being the smoothing factor determining the proclivity of the algorithm to smooth out minor spectral peaks. We apply the algorithm twice: a forward pass starting from the highest frequency bin located at i_h and a reverse backward pass starting from the lowest frequency bin. Having estimated the spectral envelope, we can now extract the source component by following $S(\omega) = \frac{Y(\omega)}{V(\omega)}$.

2.2.2. Warping function

VTLP is typically applied by remapping the center frequencies of the filterbanks in the Mel-spectrogram representation [13]. However, we do not want to lose spectral resolution inherit due to the subsampling induced by applying Mel-filterbanks on the spectrum. Subsequently, we employ the warping function directly on the extracted source and filter spectrum of the signal. The warped value F'_i of the source or filter component at frequency bin i is defined as following:

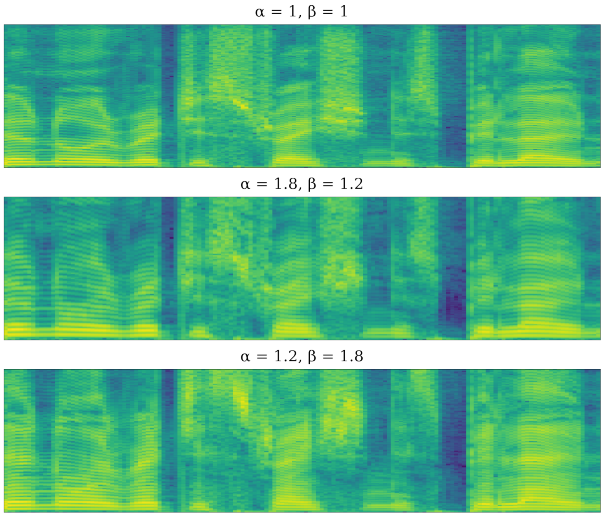

Figure 1: *Figures of an adult spectrogram (top) augmented by SFW using a high source warping coefficient α (middle) resulting in widely spaced harmonics and a high filter warping coefficient β (bottom) which mainly alters the formant locations.*

$$F'_i = F_{\lfloor \frac{1}{\lambda} i \rfloor}(1 - (\frac{1}{\lambda} i) \bmod 1) + F_{\lfloor \frac{1}{\lambda} i \rfloor + 1}((\frac{1}{\lambda} i) \bmod 1) \tag{3}$$

with λ indicating the warping coefficient. With $\lambda < 1$ and $\lfloor \frac{1}{\lambda} i \rfloor > i_h$ for $F_{\lfloor \frac{1}{\lambda} i \rfloor}$ on the right-hand side of Equation 3, the average power of the 2% upper frequency bins is used for the resulting F'_i. After applying the warping functions, we can reconstruct the augmented spectrogram by multiplying the source and filter component. Figure 1 illustrates the effect of source-filter warping with varying values for the source warping coefficient α and filter warping coefficient β.

3. Acoustic modelling

Traditional ASR systems are based on the DNN-HMM model [24, 25]. The neural network serves as the acoustic model and estimates the posterior probabilities of the acoustic units in the framed speech signal and is usually implemented as a CNN or TDNN. DNN-HMM models are still popular in low-resource ASR conditions such as children's speech recognition [5, 11].

3.1. End-to-end ASR systems

Recently, attention-based end-to-end speech recognition systems with an encoder-decoder architecture have gained state-of-the-art results on adult ASR benchmarks [3, 4]. However, training randomly initialized end-to-end models is known to require a significant amount of training data to model the latent acoustic representations robustly [26]. This poses an important limitation on the direct usage of these models in children's ASR applications.

In the context of low-resource speech recognition, promising results are recently gained by applying transfer learning techniques to adapt a model trained on large speech corpora to perform robust speech recognition on resource constrained out-of-domain data [7, 8]. We apply transfer learning on a Transformer model pre-trained on adult speech data with a masking objective to build a robust children's speech ASR system.

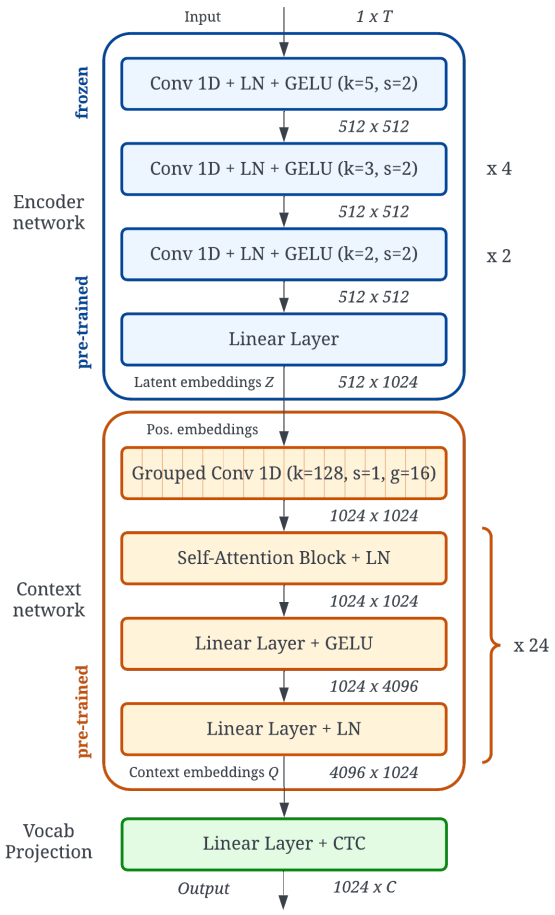

Figure 2: *Diagram depicting the network architecture for our ASR system. We denote k, s and g for kernel size, stride and group size in the convolutional layers, respectively. T refers to the temporal dimension of the input waveform and C denotes the output dimension matching the vocabulary size.*

3.2. Proposed Transformer ASR system

The architecture of our ASR system is based on the recently proposed XLS-R model [10], a Transformer architecture based on wav2vec 2.0 [8]. The model consists of two components: the *encoder network* and the *context network*. The encoder network consists of stacked blocks of temporal convolutions followed by layer normalization (LN) [27] and the GELU [28] activation function, converting the raw input waveform into a sequence of latent speech representations Z. Positional embeddings are added to the speech representations for the context network to be able to model long-range dependencies of the input signal. The positional embeddings are learned by a grouped convolution, providing the following layers relative positional information [29]. The speech representations are then used as input to the context network. The context network consists of stacked Transformer blocks, modelling contextualized acoustic representations Q.

Following [10], the model is pre-trained on cross-lingual unlabelled adult speech corpora in a self-supervised manner using a contrastive loss function in which the model needs to predict quantized audio representations from masked output latent representations of the encoder network. The learned context representations should be able to capture robust acoustic and linguistic characteristics of the input utterance. We choose to pre-train on cross-lingual adult data to make the context representations and subsequent ASR system robust against children's speech in a variety of languages without the need to pre-train and optimize separate models in future work.

After pre-training, a linear layer is added to the context network which projects the context representations Q to the vocabulary of the ASR task. We fine-tune the model using a combination of augmented adult speech with the proposed SFW strategy described in Section 2.2 and in-domain children's speech. By fine-tuning the pre-trained adult model on children's speech and augmented adult data, the network should be able to make the learned adult acoustic representations of the context network robust against the corresponding children's speech. The optimization of the model on the ASR task is done using the Connectionist Temporal Classification (CTC) [30] objective function. During this fine-tuning stage, we freeze the layers of the encoder network to prevent overfitting and reduce computational complexity. The final architecture is shown in Figure 2.

4. Experimental evaluation

To analyse the impact of the proposed SFW and transfer learning strategy for children's speech ASR, we evaluate our approach on the test set of the *PF-STAR British English Children's Speech* corpus [31]. The dataset contains 7.4 and 5.8 hours of transcribed audio for the training and test partitions, respectively. The age of the children in the dataset range from 4 to 14 years. Following other papers [32, 15, 33], we use the training subset of the WSJCAM0 corpus [34] as out-of-domain adult data, containing 15.5 hours of British English speech across 92 speakers.

4.1. Source-filter warping

To apply our proposed SFW augmentation strategy, power spectrograms with a window size of 25 ms and hop length of 10 ms are generated from the adult speech waveforms using an FFT length of 512. Subsequently, we use the algorithm described in Section 2.2.1 with $\gamma = 0.2$ to estimate the filter and source component on which we apply our warping function given by Equation 3. As we warp both components independently, we allow warping coefficients α and β to be relatively high by randomly choosing a value between 1 and 1.3.

Since our end-to-end ASR system acts on the waveform of the input signal, we need to convert the spectral representation back to the temporal domain. The Griffin-Lim algorithm [35, 36] is used to estimate the phase component of the STFT power spectrograms. To limit the computational impact, the alternating forward and inverse STFT step in the algorithm is only repeated 8 times.

4.2. Acoustic model training

Our model is pre-trained self-supervised on unlabelled adult speech data to learn meaningful acoustic representations of speech using the architecture depicted in Figure 2. More details about the pre-training step are described in [10].

During fine-tuning, we pool the training data of the PF-STAR and WSJCAM0 datasets together with equal sampling probability for both domains during batch construction. A batch size of 48 is used and the model is trained for 60K steps using the AdamW [37] optimizer. The learning rate starts at 5e-5, followed by a warmup stage of 500 steps to 1e-4 and then lin-

early decreases to 0. The input waveform is mean and variance normalized and randomly cropped according to a random start and end timestamp of the transcriptions. The crop size is limited to contain between 2 and 4 seconds of audio with no usage of SpecAugment [38]. We found this to be more effective and faster to train as opposed to using longer utterances with SpecAugment enabled. A bi-gram in-domain language model (LM) was employed to decode the test utterances, similar to [15]. The out-of-vocabulary (OOV) rate of the LM is 2.27% with a perplexity of 70.3 with respect to the PF-STAR test set.

5. Results

Table 1 shows the performance of the proposed transfer learning approach and SFW augmentation strategy. The baseline performance of our ASR system using a language model and no data augmentation strategy trained on children's data gathers a strong baseline result of 6.89% WER on the PF-STAR test set. Including the out-of-domain adult WSJCAM0 dataset improves performance on the children's test set only negligibly. We suspect this is mainly due to the extensive self-supervised pre-training of the model on adult data. However, employing the SFW strategy during training on the adult dataset, we see a relative improvement of the WER on the children's test set of 28.4% over the system without adult data augmentation. This shows that the proposed SFW strategy successfully induces the characteristics of children's speech into the adult utterances, closing the domain gap between the adult and children's speech datasets significantly. To the best of our knowledge, this is the best published result on the PF-STAR test set.

Table 1: *WER performance on the PF-STAR test set.*

Training data	**WER(%)**	
	CTC	**with LM**
WSJCAM0	40.94	18.64
PF-STAR	9.53	6.89
PF-STAR + WSJCAM0	9.48	6.79
PF-STAR + WSJCAM0 (SFW)	**6.57**	**4.86**

A performance analysis of the proposed SFW in comparison to VTLP can be found in Table 2. The baseline system is trained on adult and children's speech without any augmentation strategy. Notably, the artefacts introduced by the temporal reconstruction of the input signal from the FFT spectrum with the Griffin-Lim (GL) algorithm has a beneficial robustness effect, as shown by the *GL* experiment where we did not apply any warping augmentation but did convert the input utterances between the spectral and time domain.

The best performing VTLP configuration with a random warping factor between 1 and 1.2 improves the WER with 17.8% relative over the baseline with usage of a language model. The best SFW strategy is gained with α and β randomly varying independently between 1 and 1.3 and improves the result further with 12.9% WER relatively over the best VTLP configuration. Due to the independent warping of the source and filter component, the best SFW configuration allows for higher maximum warping coefficients as opposed to VTLP, which only has one parameter to control the warping.

Figure 3 shows the WER of each age group in the PF-STAR test set from the baseline system together with the best performing VTLP and SFW configuration from Table 2. We see

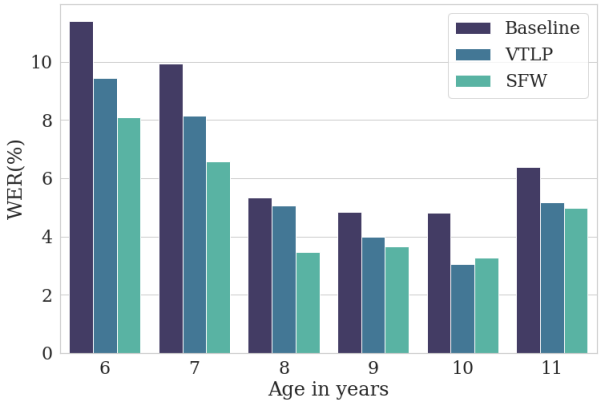

Figure 3: *Bar chart showing the WER on the PF-STAR test set for each age group using VTLP and the proposed SFW.*

Table 2: *Performance analysis of source-filter warping.*

Method	**Warp Factors**	**WER(%)**	
		CTC	**with LM**
baseline	/	9.48	6.79
GL	/	8.38	6.27
VTLP	$\eta = [1, 1.15]$	7.79	5.62
	$\eta = [1, 1.20]$	7.75	5.58
	$\eta = [1, 1.25]$	8.07	5.96
	$\eta = [1, 1.30]$	8.32	6.22
SFW	$\alpha, \beta = [1, 1.15]$	7.19	5.32
	$\alpha, \beta = [1, 1.20]$	6.92	5.07
	$\alpha, \beta = [1, 1.25]$	6.69	4.95
	$\alpha, \beta = [1, 1.30]$	**6.57**	**4.86**

that the performance increase relative to the baseline model is consistent across all ages, indicating that the proposed SFW is able to model characteristics of varying age groups. Transcription quality is clearly correlated to the age group with the WER being inversely proportional to the speaker age. An exception is the group of age 11, we suspect the degradation is mainly due to the more complex text transcriptions appearing in this age group [31]. Interestingly, we see that the largest performance improvement of SFW over VTLP is manifested in the younger age groups. We believe this is due to the benefit of independently warping the source and filter component in SFW, as younger children exhibit more varying formant locations and fundamental frequencies as opposed to older children.

6. Conclusion

In this paper, we applied transfer learning on a Transformer model pre-trained with adult speech and proposed the source-filter warping data augmentation strategy for robust children's speech ASR. Using a few hours of in-domain children's speech data, our fine-tuned Transformer model scores a WER of 6.79% on the PF-STAR children's speech test set. Applying our proposed source-filter warping strategy to close the adult and children's speech domain gap improves this strong baseline system with a final WER of 4.86%, significantly outperforming previous state-of-the-art results on the PF-STAR test set.

7. References

[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[2] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020, pp. 4218–4222.

[3] 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," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[4] C. Liu, F. Zhang, D. Le, S. Kim, Y. Saraf, and G. Zweig, "Improving RNN transducer based ASR with auxiliary tasks," in *IEEE Spoken Language Technology Workshop, SLT 2021*, 2021, pp. 172–179.

[5] R. Gretter, M. Matassoni, D. Falavigna, K. Evanini, and C. W. Leong, "Overview of the Interspeech TLT2020 Shared Task on ASR for Non-Native Children's Speech," in *Proc. Interspeech 2020*, 2020, pp. 245–249.

[6] 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," *CoRR*, vol. abs/2011.06724, 2020.

[7] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[8] 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*, vol. 33, 2020, pp. 12 449–12 460.

[9] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition." *CoRR*, vol. abs/2010.10504, 2020.

[10] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[11] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, and L. Wang, "ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech," in *Proc. Interspeech 2021*, 2021, pp. 3845–3849.

[12] G. Xu, S. Yang, L. Ma, C. Li, and Z. Wu, "The TAL System for the INTERSPEECH2021 Shared Task on Automatic Speech Recognition for Non-Native Childrens Speech," in *Proc. Interspeech 2021*, 2021, pp. 1294–1298.

[13] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[14] J. Wang, Y. Zhu, R. Fan, W. Chu, and A. Alwan, "Low Resource German ASR with Untranscribed Data Spoken by Non-Native Children — INTERSPEECH 2021 Shared Task SPAPL System," in *Proc. Interspeech 2021*, 2021, pp. 1279–1283.

[15] 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*, 2020, pp. 4382–4386.

[16] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, "Data augmentation using cyclegan for end-to-end children asr," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 511–515.

[17] G. Fant, "The source filter concept in voice production," *STL-QPSR*, vol. 1, no. 1981, pp. 21–37, 1981.

[18] H. Vorperian and R. Kent, "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data," *Journal of speech, language, and hearing research : JSLHR*, vol. 50, pp. 1510–45, 2008.

[19] R. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *Journal of speech and hearing research*, vol. 19, pp. 421–47, 1976.

[20] B. Story and K. Bunton, "Formant measurement in children's speech based on spectral filtering," *Speech Communication*, vol. 76, 12 2015.

[21] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[22] A. V. Oppenheim and R. W. Schafer, "Digital signal processing," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, pp. 146–146, 1978.

[23] J. D. Markel and A. H. Gray, *Linear prediction of speech*. Springer-Verlag Berlin ; New York, 1976.

[24] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[25] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proc. Interspeech 2016*, 2016, pp. 2751–2755.

[26] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, Battenberg *et al.*, "Deep speech 2 : End-to-end speech recognition in english and mandarin," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48, 2016, pp. 173–182.

[27] J. Ba, J. Kiros, and G. Hinton, "Layer normalization," *NIPS 2016 Deep Learning Symposium*, 2016.

[28] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[29] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for ASR," *CoRR*, vol. abs/1904.11660, 2019.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks," vol. 2006, 2006, pp. 369–376.

[31] 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," in *Proc. Interspeech 2005*, 2005, pp. 2761–2764.

[32] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation," in *Proc. Interspeech 2016*, 2016, pp. 1598–1602.

[33] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "A formant modification method for improved asr of children's speech," *Speech Communication*, vol. 136, pp. 98–106, 2022.

[34] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1995, pp. 81–84.

[35] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," in *ICASSP*, 1983.

[36] N. Perraudin, P. Balazs, and P. Søndergaard, "A fast griffin–lim algorithm," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[38] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.