Proceedings Paper:
Nomo Sudro, P., Ragni, A. and Hain, T. (2023) Adapting pretrained models for adult to child voice conversion. In: 2023 31st European Signal Processing Conference (EUSIPCO) Proceedings. 2023 31st European Signal Processing Conference (EUSIPCO), 04-08 Sep 2023, Helsinki, Finland. Institute of Electrical and Electronics Engineers (IEEE), pp. 271-275. ISBN: 9789464593600. ISSN: 2219-5491. EISSN: 2076-1465.

https://doi.org/10.23919/EUSIPCO58844.2023.10289993

Adapting pretrained models for adult to child voice conversion

Protima Nomo Sudro, Anton Ragni, Thomas Hain
Department of Computer Science
Speech and Hearing Research Group, University of Sheffield, UK
{p.nomo.sudro, a.ragni, t.hain}@sheffield.ac.uk

***Abstract*—Due to widespread lack of parallel data for adult to child voice conversion (VC), non parallel VC techniques have grown in popularity. Methods, such as encoder-decoder model, have achieved good performance in adult-to-adult VC. It provides flexibility by either training each module separately or exploit pretrained models. These pretrained models are only available for adult speech. In case of children speech, we do not have enough data to train all the modules of a robust encoder-decoder based VC system. In a limited data scenario, we can only train the decoder module using target speech. Specifically, we find that adult to child VC using a pretrained encoder and trained decoder with child speech does not yield spectral variability of a child speech. The reason being gross spectral mismatch between adult and child speech. We address this mismatch by exploiting a warping mechanism to modify the acoustic attributes based on child speech. We conduct objective and subjective evaluations on CMU and CSLU kids corpus and one adult actress data. Results show that the proposed method reduces MCD and F0 RMSE by 0.67 and 0.03 respectively. For subjective evaluations we observe a relative MOS improvement of 10.7% for naturalness and 18.23% for similarity.**

***Index Terms*—Child speech, adult speech, voice conversion, encoder-decoder model**

I. Introduction

A voice conversion (VC) system takes speech from a source speaker and generates an output speech that sounds like a target speaker [1]. While performing the conversion, such systems are expected to convey the same linguistic content from source to target speakers whilst converting pitch and intonation patterns, prosody, and other spectral attributes to match those of target speakers [2]. VC techniques can be broadly classified as parallel and non parallel methods based on the availability of training data. Parallel VC techniques usually offer high quality. However, it is not always possible to collect sufficient quantities of parallel data [3]. To reduce reliance on parallel data, recent VC studies have explored leveraging large amounts of non parallel data and feature combination approaches. Among the many non parallel VC techniques proposed to date, an encoder-decoder based VC has become popular [4]. This VC technique consists of a feature extractor module, synthesis module, and a vocoder. Each of these systems can be trained separately, which offers additional flexibility.

This work is supported by KNOWLEDGE TRANSFER PARTNERSHIP between The University of Sheffield and Zoo Digital Group plc [Ref No KTP 12423]

Voice conversion techniques have been applied for various application such as customizing audiobook and avatar voices, computer-assisted pronunciation training, speech to singing conversion, speech synthesis, communication aid for speakers with impaired speech, speech enhancement, and dubbing [5]–[7]. In the current work, we focus on adult to child VC for dubbing. The process of dubbing involves translating original dialogue from media based on the script, tone, genre, emotions and synchronising them with lip movements. Dubbing child speech, however, is difficult because of the limited voice resources, regulations associated with child casting, expression of desired tone and mood during recording [8]. To address this issue by means of VC, such as the encoder-decoder technique, we would need some quantities of acted speech data from professional voice talents. The acted speech data is rare and quite different compared to read speech data used in typical VC studies. Furthermore, the physiological attributes of a child presents acoustic variability such as, pitch and formants which affects overall spectral and temporal characteristics [9]. Moreover, compared to adult speech, child speech exhibit different characteristics such as, low speaking rate, pronunciation problems, false starts, disfluencies, and different non-speech sounds [10].

The limited training data scenario motivates the use of encoder-decoder based VC. The specific technique examined in this study consists of a bottle-neck feature extractor and a sequence-to-sequence (seq2seq) synthesis module. A pre-trained bottle-neck feature extractor was used in this study, which is trained on 960 hours of LibriSpeech data [11]. The synthesis module was trained using CMU and CSLU child speech corpus [12], [13]. The conversion based on encoder-decoder model has a good quality speech, but the converted speech has less similarity with the target speech because of higher formant frequency and other spectral variability. The analysis of converted speech revealed the need for spectral transformation to match characteristics of child speech more closely. To achieve this, we exploit the use of a warping technique as a post processing method. Similar techniques have been previously investigated for mapping source to target speech characteristics in VC, speech synthesis, and automatic speech recognition (ASR) studies [14]–[16] but not particularly for adult to child VC.

To the best of our knowledge, only few adult to child VC studies have been reported in the literature. One of the

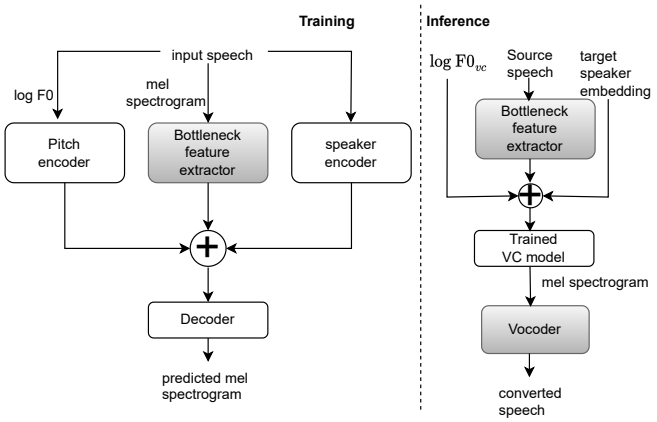

Fig. 1. Illustration of the training and inference stage of the voice conversion system. Grey colored boxes denote publicly available pretrained models.

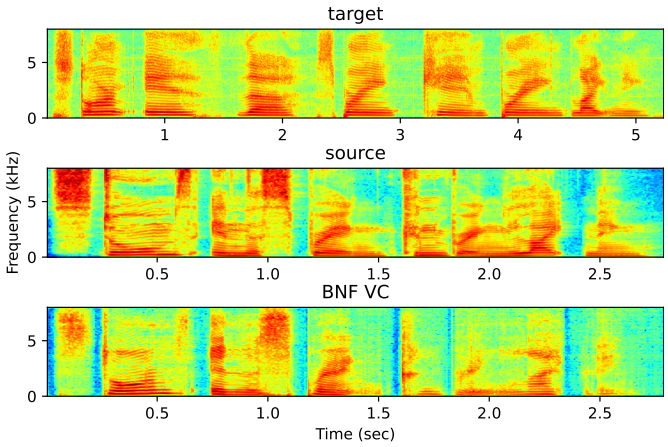

Fig. 2. Illustration of spectrogram for an utterance 'storms in the spring can bring lightning' (a) target speech, (b) source speech, and (c) converted source speech using BNF VC

studies reported using read speech for training adult to child CycleGAN VC model for ASR application [17]. Other studies by [18], [19] reported using Gaussian mixture model (GMM) based adult to child VC for speaker adaptation and dubbing. Our work is different from all of these in several important aspects: (a) first, our work considered using real media data from professional voice talents, (b) next, in [18], dubbing was performed for Indonesian language using several words only, (c) in [17], VC was performed for data augmentation and did not investigate target speaker quality and similarity, (d) and finally, in [19] VC was applied to the output of a speech synthesizer.

The remaining paper is organized as follows. The proposed VC framework is discussed in Section II. The experiment details are reported in Section III. Section IV includes results and discussions. Finally, the work is concluded in Section V.

II. Methods

This section explains the transformation of an adult speech into that of a child speech based on warping in combination with encoder-decoder based VC framework.

A. Encoder-decoder VC model

The VC approach followed in this work is shown in Figure 1, the encoder consists of a bottle-neck feature extractor (BNF) and a pitch encoder. The decoder is a seq2seq synthesis module [11]. Hereafter, we refer to this part as a BNF VC.

Voiced and unvoiced regions are first extracted prior to training. Next, the pretrained BNF extracts content representations from mel-spectrograms. The log F0 are converted from source to target F0 using logarithm Gaussian normalised transformation (log F0$_{vc}$). Together with the BNF and unvoiced-voiced (UV) flags they are added element-wise, and concatenated with a target speaker embedding vector to form encoder outputs. Speaker embedding vectors are obtained from a speaker encoder model, which is trained before training the VC module. The decoder has a network structure similar to Tacotron 2 except mixture of logistics (MoL) attention. MoL refers to the computation of a set of distribution parameters which corresponds to mixture coefficient, mean and scales [20].

During inference stage, mel-spectrogram, log F0 and UV flags are extracted from a source utterance. Further, bottle neck features extracted from BNFs, log F0$_{vc}$, and target speaker embedding vectors are input to the trained VC model (i.e, decoder). The predicted mel-spectrograms from the decoder are input to the pretrained Hi-Fi GAN vocoder to generate the output speech (BNF VC).

Figure 2 shows spectrograms of a child (target), adult (source), and BNF VC converted speech. The spectrograms depict higher formant frequencies and low speaking rate of the child speech compared to the adult speech. In line with the literature, it is observed that source and target speech exhibits a pitch of 203 Hz and 294 Hz respectively [9], [21]. The corresponding average pitch value for converted speech is 265 Hz respectively. Although, the converted utterance pitch is close to the target speech, but the formant locations does not match with the target speech. For example, from Figure 2, we can see this difference for formants around 4 kHz for converted speech (BNF VC) and that of target speech. It is also to be noted that formants above 4 kHz have lower energy distribution in the converted speech compared to the target speech. This suggests that using pitch based warping could yield a transformation that reallocates the formant frequencies. Further, an energy scaling factor can be exploited to modify the the low energy distribution.

B. Proposed approach

To address the spectral variability issues, we explored warping technique. Warping is a technique commonly applied for generating augmented data in child ASR studies [22]. Due to the lack of data, many child ASR systems are trained using adult speech data. Motivated by human perception studies, various works have reported on the use of F0 in the frequency warping technique [16] which would act as a normalising factor for the speech frequency spectrum [16], [23]. Hence, we could exploit the same warping technique in combination with non parallel VC technique to increase similarity between target child speech and converted speech.

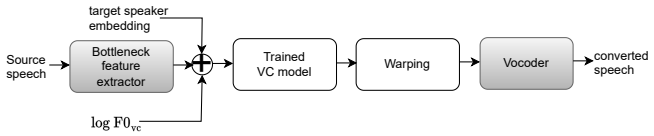

Fig. 3. Illustration of the proposed voice conversion system, grey colored box indicate publicly available pretrained models

The warping is performed in the mel domain by normalising speech spectrum on a per-frame basis using the per-frame estimate of the F0. As shown in Figure 3, the warping can be applied only during the inference stage. The pitch based warping is applied to every frequency components of the frame in the utterance using the following equation,

$$f^k_{\text{warp}} = f^k_{\text{source}} - (F0_{\text{source}} - F0_{\text{target}}) \quad (1)$$

here, f^k_{warp}, f^k_{source} denotes warped frequency and the frequency to be warped from the source spectrum. Here k ranges from $1 \cdots N$, N corresponds to the total number of frequency points expanding the bandwidth from 20 Hz to 8 kHz. $F0_{\text{source}}$ and $F0_{\text{target}}$ corresponds to the median F0 of the source and target utterance respectively. Using equation 1, the entire speech spectrum is normalized based on the F0. Basically, f_{warp} maps the representative frequencies of the source spectrum onto that of the target spectrum. f_{warp} varies with F0 values per frame of the source utterance.

Once the warped frequency is obtained, the energy distribution of the spectrum is modified as, $S^k_e = \beta \times S^k_w$, S^k_w and S^k_e denotes the warped and corresponding energy modified spectrum. The scaling factor β is empirically obtained based on the energy distribution of the target speech. Therefore, the scaling factor varies for each of the target speaker.

In literature warping alone is applied to transform adult to child speech spectrum. However, standalone warping applied for data augmentation or improving ASR systems may not always resemble pitch and speaker similarity with a target child.

III. Experimental setup

A. Data

The datasets used in this work are available in the public domain. The CMU child speech corpus [13] is used to train the voice conversion model. Each of 76 recruited speakers (52 female and 24 male) were expected to utter 356 unique texts. However, fewer utterances are available per each child with the remaining ones either not recorded or filtered out. The total number of available utterances is 5191 (9.1 hours). The age of the child speakers ranges between 6 and 11 years. The data is read speech sampled at 16 kHz. In addition to the CMU speech corpus, we also used the CSLU child corpus [12]. It consists of 100 speakers per educational grade level (kindergarten to the 10^{th} grade) uttering 208 isolated words. We have used 40947 utterances (43.15 hours) of read speech.

In this work, 80% of the utterances are used for training and 10% each are used for validation and testing respectively. In addition to the child speech corpora, a parallel adult speaker data is recorded. The recording is done in ZOO Digital studio, London. The adult speaker (a female actress) is chosen based on the accent similarity to that of the child speakers (i.e., American accented speech). The parallel adult speech recording consists of (a) 356 utterances from CMU corpus and (b) 208 isolated words from CSLU corpus.

B. Implementation details

Prior to training VC module, the speech signals are down-sampled to 16 kHz, split into frames of 50 ms using 10 ms shift and multiplied with the Hanning window. The Mel-spectrograms, UV flags, and F0 are extracted using the WORLD analysis system [24]. The dimensionality of all the spectral feature vector is 80. However, the dimension of the speaker encoder is 40 and window size of 25 ms and 10 ms hop size.

The VC module is trained using the continuously interpolated F0 (log F0), UV flags, and BNFs. The latter are obtained by feeding 80-dimensional mel-spectrograms into the pretrained BNF. Prior to feeding the input features to the BNF block, an utterance-level mean and variance normalisation is performed. The BNFs from the BNF are passed through the BNF prenet. The BNF prenet contains two bidirectional GRU layers with 256 hidden units per direction. The pitch encoder consists of convolution network structure and it takes log F0 and UV flag features as the input. Log F0 and UV are extracted using the same frame shift as that of mel-spectrograms. Since, BNFs only have a quarter of frames due to the down-sampling by a factor of 4 in the feature extraction process, log F0s and UVs are also down-sampled by a factor of 4 to match the same time resolution. This is performed by using 1-dimension convolution layer with a stride of 2 and a hidden dimension of 256. Possible speaker information is removed by adding an instance normalisation layer without an affine transformation after each of the convolutional layers of the pitch encoder.

While performing conversion, the output of the pitch encoder and the BNF prenet are added element-wise alongwith speaker vectors. Later, the decoder is fed with encoder outputs to generate predicted mel-spectrograms. Then, the transformed mel-spectrograms obtained after warping is used to synthesize the waveform.

IV. Results and discussion

Prior to describing objective and subjective evaluations performed in this work, we would like to illustrate the characteristic differences between adult and child speech and the relative impact of BNF VC and BNF VC combined with warping (BNF + warp VC) on the modified adult speech. For this purpose, the corresponding spectrograms are shown in Figure 4. In Figure 4 we observe that after the BNF VC the pitch range has changed from 203 Hz to 265 Hz. However, the formant frequencies and energy of high frequency components can be observed to be in the range of an adult speech. These are the important factors that differentiates the perception of an adult from that of a child speech [16]. Therefore, these issues necessitate the importance of warping based on the

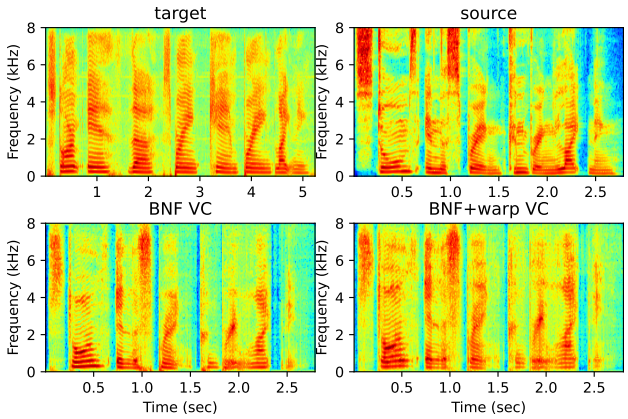

Fig. 4. Spectrograms of the sentence 'storms in the spring can bring lightning' for : target, source, BNF VC, and BNF + warp VC

pitch, or other techniques, for enabling child like speech perception. The impact of BNF + warp VC shows higher formant frequencies compared to the BNF VC transformed speech and original adult speech. Additionally, the low energy distribution in the higher frequency region (specially above 4 kHz) needs to be improved close to the target speech. To compensate the energy mismatch, a scaling factor is used to modify the energy of high frequency components.

A. Objective evaluation

The objective metrics used in this study are Mel-cepstrum distortion (MCD), F0 root mean square error (F0 RMSE), character/word error rate (CER/WER) metrics.

The MCD measures spectral distortion between adult and child Mel cepstral coefficients and is obtained by computing the mean of MCD values across all the frames. The F0 RMSE compares the accuracy of F0 conversion. The overall F0 RMSE value is obtained by computing the mean of F0 values across all the voiced frames. The CER and WER are computed by comparing texts to predictions produced by a transformer based ASR system transcribing speech (transformed and target) [25]. The ASR was build using adult speech.

As an initial investigation, we explored two baseline VC approaches, BNF VC and VCC2020, on two datasets, CMU and CSLU. VCC2020 is an abbreviated form for the voice conversion challenge 2020 recipe [26]. From Table I it is observed that BNF VC shows improved performance compared to the VCC2020 baseline. Table I also shows that when the BNF VC approach is applied, the F0 RMSE changes only slightly. However, the F0 RMSE and other metrics showed significant reduction when BNF + warp VC is used. This implies that the difference between converted speech and target child speech characteristics is reduced compared to the BNF VC alone. For comparison purpose, a BNF VC model trained using adult speech, particularly, VCTK corpus and CMU ARCTIC database released by authors in [11] is used to observe the performance on adult to child VC. We denote this model as BNF_aVC. We observe a similar trend for BNF_aVC which exhibits improved performance with respect to the CER and WER values but MCD and F0 RMSE are not significantly

TABLE I
PERFORMANCE OF DIFFERENT VC APPROACHES USING DIFFERENT EVALUATION METRICS ON CMU AND CSLU CORPUS, SRC AND TGT DENOTES SOURCE AND TARGET

Dataset	VC method	MCD	F0_{RMSE}	CER	WER
CMU	None (src vs tgt)	9.08	0.73	22.4	38.4
	VCC2020	8.89	0.65	30.0	37.8
	BNF VC	8.14	0.57	20.5	31.0
	BNF+warp VC	7.47	0.54	18.5	28.7
VCTK + CMU ARCTIC	BNF_aVC	7.76	0.74	14.3	21.3
	BNF_a + warp VC	7.42	0.68	13.5	19.9
CSLU	None (src vs tgt)	9.70	0.70	22.0	35.6
	VCC2020	8.82	0.69	21.5	35.2
	BNF VC	9.12	0.68	21.2	32.9
	BNF+warp VC	8.09	0.64	20.1	30.9

improved compared to the BNF VC model trained on child speech corpus. The lower CER and WER of the BNF_aVC are attributed to the fact that a large amount of adult speech data is used in the training. In case of the BNF_aVC, we also observe that when warping is applied in combination with the VC approach, the performance is improved.

B. Subjective evaluation

The subjective evaluation is conducted to assess both naturalness and preference when comparing converted speech to target speech. A total of 11 native English speakers have participated in the study. Each of them were provided with 10 sets of utterances. Each set comprised of speech from the source, the target, the BNF VC model, the BNF + warp VC model, the BNF_aVC model, and the BNF_a + warp VC model. To investigate the impact of warping, we focus on the CMU corpus as it contains more sentences compared to the CSLU corpus. The study was conducted in a quiet room using the same headphone and computer. All the tests were performed in the same setting for all the participants.

In the first subjective evaluation study, the listeners were asked to rate the perceptual quality of the speech using a 5-point Likert scale, where $1 =$ bad, $2 =$ poor, $3 =$ fair, $4 =$ good, $5 =$ excellent. In the mean opinion score (MOS) test, random samples were generated from the 4 comparative approaches and presented to the listeners. The average scores across all the listeners are computed to obtain the reported MOS scores. The mean and standard deviation are shown in Table II. The MOS score of approximately 3 indicates that the perceptual quality of the converted speech is fair. Compared to the BNF VC approach, the proposed warping technique shows better results.

The converted speech was also evaluated in terms of subjective perceptibility by comparing similarity of the transformed

TABLE II
NATURALNESS AND PREFERENCE TEST ON SIMILARITY FOR TARGET AND BNF VC AND BNF+WARP VC OUTPUT SPEECH

VC method	Naturalness $\mu \pm \sigma$	Similarity Set1	Similarity Set2
BNF VC	2.99 ± 1.22	1.79	3.29
BNF + warp VC	3.31 ± 1.03	1.25	3.89
BNF_aVC	3.25 ±1.13	3.99	1.02
BNF_a + warp VC	3.35 ±1.05	3.88	1.18

speech to the adult and child speech. In this test also, a 5-point Likert scale is used for rating. For each VC method, two similarity tests were conducted: (a) Set 1 measures similarity between the converted speech and the adult speech, and (b) Set 2 measures similarity between the converted speech and the child speech. The results are shown in Table II. It can be observed that the Set 2 exhibit higher similarity scores with the child speech as compared to the lower similarity score with the adult speech for BNF VC trained with CMU speech. This implies that the BNF VC approach trained using CMU speech corpus results in close similarity with child speech characteristics and with warping, we can see further improved results. For BNF_aVC model, significant similarity with the adult speech is observed from Set 1. The is attributed to the fact that the BNF_aVC model is trained using large amount of adult speech and hence it shows lower similarity with child speech. For reference, some of the audio samples can be found in https://drive.google.com/drive/folders/1iKcXfcFazs24IMTcLnenAxog5V8gNc6A?usp=sharing

V. Conclusions

In this work, we conducted a study on the use of encoder-decoder based non parallel VC approach (BNF VC) for adult to child VC. Through the objective and subjective evaluation we observed that the BNF VC approach when trained using child speech corpus shows significant improvement in terms of pitch and similarity compared to the previous work. Yet many important features linked with child speech remain poorly handled by the BNF VC approach. To address some of them this paper proposed a simple, efficient and effective warping technique. Combined with the BNF VC approach this technique shows further improvement in the quality and similarity of adult to child VC. In addition to more elaborate schemes that could better account for child speech properties, future work will examine speaking rate transformation and more careful data preparation strategies that could account for known artefacts of examined child speech corpora (pauses, laugh, etc.). In addition, the BNF VC output might be affected due to the pretrained vocoder that was trained using adult speech. Therefore, vocoder for children speech is worth exploring.

References

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] D. Rentzos, S. Vaseghi, Q. Yan, and C.-H. Ho, "Voice conversion through transformation of spectral and intonation features," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–21.

[3] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[4] Z. Yi, W. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020 — intra-lingual semi-parallel and cross-lingual voice conversion —," in *Proceedings of Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, pp. 80–98.

[5] T. L. New, M. Dong, P. Chan, X. Wang, B. Ma, and H. Li, "Voice conversion: From spoken vowels to singing vowels," in *2010 IEEE International Conference on Multimedia and Expo*, 2010, pp. 1421–1426.

[6] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[7] C. Veaux, J. Yamagishi, and S. King, "Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction," in *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, 2013, pp. 107–111.

[8] S. Turunen *et al.*, "Interplay of verbal and visual: Concretisation as a dubbing translation strategy in children's tv show kit'n'kate," 2017.

[9] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[10] R. Sinha and S. Shahnawazuddin, "Assessment of pitch-adaptive front-end signal processing for children's speech recognition," *Computer Speech & Language*, vol. 48, pp. 103–121, 2018.

[11] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[12] K. Shobaki, J.-P. Hosom, and R. Cole, "The ogi kids' speech corpus and recognizers," in *Proc. of ICSLP*, 2000, pp. 564–567.

[13] M. Eskenazi, J. Mostow, and D. Graff, "The cmu kids speech corpus (ldc97s63)," *Linguistic Data Consortium (http://www. ldc. upenn. edu), University of Pennsylvania (viewed 8-27-07)*, 1997.

[14] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[15] C. Valentini-Botinhao, Z. Wu, and S. King, "Towards minimum perceptual error training for dnn-based speech synthesis," in *Sixteenth annual conference of the international speech communication association*, 2015.

[16] G. Yeung and A. Alwan, "A frequency normalization technique for kindergarten speech recognition inspired by the role of f0 in vowel perception," *Interspeech 2019*, 2019.

[17] 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 *Interspeech*, 2020, pp. 4382–4386.

[18] F. M. Mukhneri, I. Wijayanto, and S. Hadiyoso, "Voice conversion for dubbing using linear predictive coding and hidden markov model," *Journal of Southwest Jiaotong University*, vol. 55, no. 4, 2020.

[19] O. Watts, J. Yamagishi, S. King, and K. Berkling, "Synthesis of child speech with hmm adaptation and voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1005–1016, 2009.

[20] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[21] 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. 47 628–47 642, 2022.

[22] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's asr." in *Interspeech*, 2016, pp. 3459–3463.

[23] G. Yeung, R. Fan, and A. Alwan, "Fundamental frequency feature normalization and data augmentation for child speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6993–6997.

[24] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[26] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.