INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Data augmentation for children ASR and child-adult speaker classification using voice conversion methods

Zhao Shuyang, Mittul Singh, Abraham Woubie, Reima Karhila

Huawei Finland Research Center

shuyang.zhao@huawei.com, mittul.singh@huawei.com, abraham.zewoudie@huawei.com, reima.karhila@huawei.com

Abstract

Many young children prefer speech based interfaces over text, as they are relatively slow and error-prone with text input. However, children ASR can be challenging due to the lack of transcribed children speech corpora. In this paper, we investigate a voice conversion method based on WORLD vocoder to generate childlike speech for data augmentation. Since noise may lead to severe artifacts in converted speech, we also investigate using speech enhancement to improve the quality of converted speech. On a publicly available children speech corpus, we evaluated the performance of the proposed data augmentation method against existing data augmentation methods based on linear prediction coefficients. Our proposed data augmentation method substantially outperformed the prior work on children ASR. Additionally, on a task to classify the speaker, adult or child, data generated using our proposed method was shown to mimic real children better compared to the reference methods.

Index Terms: ASR, child-adult speaker classification, data augmentation, voice conversion, speech enhancement

1. Introduction

Children are an important set of voice search users. Due to children's limited typing and spelling skills, voice interfaces for search – powered by automatic speech recognition (ASR), are attractive to them [1]. However, children speech data is not publicly available in many languages. Without training on children speech, the ASR performance is typically low when tested by young users, especially for ages under 12 [2]. Thus, correctly recognizing children's voice queries is a challenging task.

Children speech is different from adult speech in many aspects, such as fundamental frequencies, formants, spectral variability, and vowel lengths. Several techniques have been proposed to address the mismatches in the acoustic characteristics of children and adults. Vocal tract length normalization (VTLN) [3] compensates the different vocal tract lengths by warping the frequency spectrum in the filterbank analysis. Spectral warping techniques [4, 5, 6] shift the formants by scaling the spectral envelope of an audio signal to approximate the distribution of formants, based on methods such as modification of linear prediction coefficients. VTLN and spectral warping can be used to simulate the formants of children. However, mismatches in other characteristics of children speech such as fundamental frequencies (F0) and harmonic structure of high F0 speech can not be handled using only VTLN or spectral envelope warping.

Voice conversion aims at generating synthetic speech that perceptually resembles children speech. This technique has been investigated using deep learning based techniques, such as CycleGAN [7, 8]. Deep-learning-based methods require children speech as training data, which is scarce in most languages. There are also voice conversion methods using signal-processing-based vocoders, such as STRAIGHT [9] and WORLD [10]. These vocoders have the potential of converting adult speech to be sufficiently childlike for data augmentation, and they do not require children speech for training. Yet, no prior work has used similar methods to improve children ASR to the best of our knowledge.

In this paper, we propose data augmentation for children ASR using a voice conversion method to generate childlike speech based on WORLD vocoder, which includes a series of signal processing algorithms [11, 12, 13] for speech analysis and synthesis. In order to map adult speech characteristics into children ones, modifications are made based on a children acoustic study [14]. The spectral envelopes are warped (Section 2.2.1); the fundamental frequencies are shifted (Section 2.2.2); the vowel lengths are stretched (Section 2.2.3). For training data recorded in real-world scenarios, background noise in the audio signal may lead to a bad estimation of the fundamental frequencies and spectral envelopes of the speech [11]: In these cases, direct voice conversion may lead to strong artifacts that cause the converted speech to deviate from the originally spoken words. Thus, we also propose using speech enhancement [15, 16] to separate speech from the background noise before performing the voice conversion (Section 2.1).

In our experiments (Section 3), the proposed data augmentation method is evaluated on two tasks. Firstly, we set up an ASR experiment to investigate the Word Error Rates for children speech, with the augmented data included in the training. Secondly, we set up an child-adult speaker classification task as an objective measurement of the resemblance between the augmented data and real children data. Finally, we present our concluding remarks in Section 4.

2. Method

We propose a method to convert adult speech to childlike speech. For simplicity, we call the process ”childrenization” in the rest of the paper. The motivation for childrenization is to improve children ASR when only adult speech corpora are available for training. An overview of the proposed childrenization method is illustrated in Figure 1. Speech enhancement is performed to reduce background noise before the childrenization process, including analysis, modification, and synthesis. The analysis and synthesis follow the framework of WORLD vocoder [10, 12], and the modification follows an acoustic analysis about the developmental changes in children speech [14]. The code and childrenized examples are available at `https://github.com/zhao-shuyang/childrenize`.

10.21437/Interspeech.2023-702

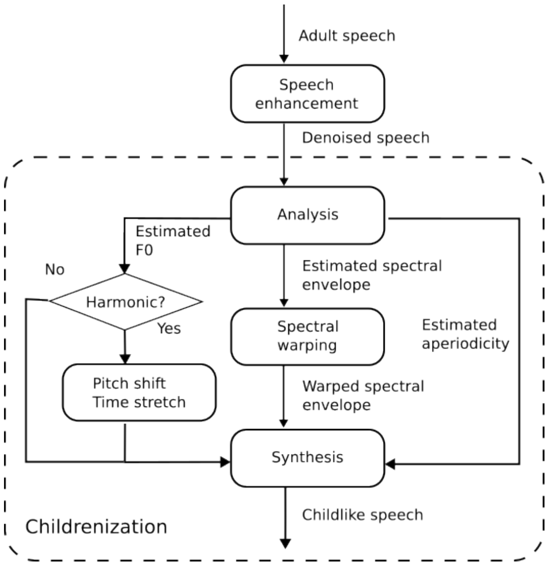

Figure 1: *An overview of the proposed childrenization method.*

2.1. Speech enhancement

The proposed childrenization algorithm relies on accurate F0 estimation. Background noise results in bad F0 estimation and thus, cause synthesized speech to be severely degraded. Degraded speech sometimes deviates from the spoken words in the original signal. To address this issue, we propose using speech enhancement to denoise speech signals before performing childrenization.

In this study, we follow the work of [16], which associates its work with a publicly available code repository [1]. It proposes a speech enancement method based on the estimation of a priori signal-to-noise ratio (SNR). Here, a temporal convolutional neural network (TCN) [17] is trained to map the magnitude spectrogram into the cumulative probabilities of SNR corresponding to each frequency bin and each time step in the spectrogram. SNR is restored from the cumulative probabilities using the quantile function based on their means and variances in each frequency bin in the training dataset. Wiener filter is used to derive the denoised speech spectrogram from the estimated SNR. For details, we refer the reader to the original work [16].

2.2. Childrenization

To match the characteristics of children speech observed in [14], we modify adult speech in three aspects: formants, pitch and vowel lengths. We aim at these speech characteristics randomized according to statistics for ages 5 to 12. The modification of each aspect is introduced in one of the following subsections.

2.2.1. Spectral warping

Perceptually, vowels are described by their formants, peaks in the spectral envelope. In order to match the formant characteristics of children, a modification is made to the spectral envelope of adult speech. The estimation of the spectral envelope is based on CheapTrick algorithm [11]. Spectral warping is performed on an estimated spectral envelope, mapping a frequency f into a new frequency f'. The proposed spectral warping function is defined based on our observations in [14].

The scaling factors of the first three formants from male adult to male children are similar, accoding to the statistics shown in [14, Figure 6]. Thus, we use a linear spectral frequency warping function for the whole frequency range as $f'_{male} = \alpha f$, where a frequency scaling factor α is randomly chosen from the range $[1.2, 1.4]$.This roughly corresponds to the formant frequency scaling factor statistics of male children from ages 5 to 12.

The formant scaling factors from female adult to female children appear lower as formant frequency increases. We propose using a piece-wise linear warping function, similar to [3, 6]. A low-cut frequency F_{low} and a high-cut frequency F_{high} divide the frequency from zero to Nyquist frequency into three pieces. Each piece has its own slope value and the piecewise spectral warping function is

$$f'_{female} \begin{cases} \beta_{low} f, & \text{if } f \leq F_{low} \\ F'_{low} + \beta_{mid}(f - F_{low}), & \text{if } F_{low} < f \leq F_{high} \\ F'_{high} + \beta_{high}(f - F_{high}), & \text{if } f > F_{high}. \end{cases} \tag{1}$$

A mid-frequency range slope β_{mid} is randomly set in the range $[1.1, 1.25]$. The low-frequency range slope β_{low} is set to the square of β_{mid}, as $\beta_{low} = \beta^2_{mid}$. As a result, the warping factor f'_{female}/f decreases from β^2_{mid} at F_{low} to approximately β_{mid} at F_{high}. Following this, the F_{low} and F_{high} are mapped to F'_{low} and F'_{high}, respectively. In the high-frequency range, the slope value is set to $\beta_{high} = ((Fs/2) - (F'_{high}))/((Fs/2) - (F_{high}))$ closing the range of warped frequency up to Nyquist frequency Fs.

We use the estimated mean F0 within an utterance to determine the gender of the speaker. We observed in [14, Figure 3a] that the gender of an adult speaker could be roughly separated using $160Hz$ as a threshold. The utterance with an estimated mean F0 above the threshold is considered to be made by a female adult, otherwise a male adult.

2.2.2. Pitch shifting

We made modifications to adult speech on the aspect of the pitch, which is quantified by the fundamental frequency (F0). The modifications are based on [14, Figure 3a], which illustrates the mean and standard deviation of the F0 from age 5 to age 18. We randomly set a target of the mean F0 for each utterance, $\bar{p}_{target}$, from the range $[240, 300]$ Hz. This value roughly corresponds to the mean F0 range within one standard deviation for the ages 5 to 12.

For each frame, the F0 is estimated using the Harvest algorithm [13], denoted as $p_{estimate}$. The estimated mean F0 of the utterance, $\bar{p}_{estimate}$, is computed excluding non-harmonic frames and frames with an estimated F0 lower than 50 Hz. An estimated F0 lower than 50 Hz is considered to be from non-speech sound sources or due to estimation error. The target F0 of a frame, p_{target} is obtained by shifting the original F0 by the difference between the target mean and the original mean: $p_{target} = p_{estimate} + (\bar{p}_{target} - \bar{p}_{estimate})$.

Most prior work, including VTLN [3] and LPC spectral warping [4, 5, 6] do not deal with pitch. We note that children's high pitch can largely affect the acoustic features of vowels. Figure 2 illustrates the log energies of 40 mel bands from a sample vowel frame corresponding to an adult male speech signal, its spectral-warped version, and its spectral-warped and pitch-shifted version. The harmonic structures in the low-pitch signals are smoothed out by the mel filters, whereas the harmonic structures are visible in the pitch-shifted version. The interval between harmonics proportionally increases with the F0. With

[1] https://github.com/anicolson/DeepXi

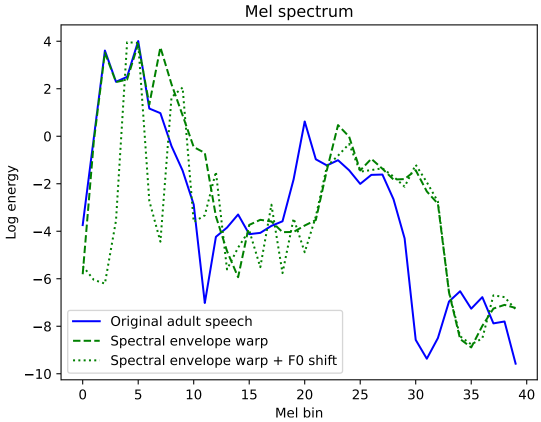

Figure 2: *Sample mel-weighted frequency bins of a vowel frame in an adult male speech signal, its spectral-warped version, and its spectral-warped and pitch-shifted version. The spectral envelope has been linearly warped by a factor of 1.3 and the F0 has been shifted from about 120 Hz to 250 Hz.*

a high F0, the intervals between harmonics can be larger than the mel filter bandwidth. As a result, the harmonic structures are not smoothed out below bin number 22 in the pitch-shifted signal.

2.2.3. Vowel length stretching

Lastly, a modification is made to the vowel lengths. According to [14, Figure 1a], the average lengths of vowels decrease from age 5 to age 12, from 280 ms to 180 ms. Based on this observation, we randomly set a vowel length stretching factor γ in the range $[1.1, 1.4]$. The change of vowel lengths can have an impact on languages that distinguish vowel phonemes by their lengths.

In the implementation, an utterance is divided into segments based on the change points of harmonicity, which is estimated using the F0 estimation algorithm [13]. The vowel length stretching factor γ is used to scale the hop length of harmonic segments during the synthesis, while the non-harmonic segments remain unspoiled.

3. Experiments

3.1. Experimental setup

3.1.1. Data

Samrómur [18, 19] is the only freely available children speech recognition corpus to the best of our knowledge, thus our evaluation is based on this Icelandic dataset. As for training data, we build our models using the freely available Icelandic Malrómur speech recognition dataset [20], which contains only adult speech. Background sound sources exist in some of the utterances for both Samrómur and Malrómur. We split the Malrómur data into train, development and test sets; and release the splits for future comparisons.

In terms of size, the Samrómur dataset contains 127.4 and 1.8 hours as the training and test sets respectively. In addition to the full Sarmómur dataset, we also take a 5.4 hours subset from Samrómur training set for experiments. The experiment simulates a situation where a small amount of children speech data is available for training. Malrómur datasets have 109.0, 13.4 and 13.6 hours in the training, development and test sets respectively. Henceforth, we refer to Samrómur as the *Real Children* corpus (RC) and Malrómur as *Real Adult* corpus (RA). The 5.4-hours subset of the Samrómur training set is denoted by RC_5.

3.1.2. Training data augmentation

As described above, the experiment involves a real adult corpus (RA) and a real children (RC) corpus. Each evaluated data augmentation method is performed on a copy of the training split of RA. The evaluated data augmentation methods are listed below.

RAL denotes RA spectral-warped based on a linear prediction coefficient (LPC) method [4], using its open-source implementation. The warping factor of RAL is set to -0.05, which is the best-performing value for children ASR in [4, Table 3]. RAC denotes childrenized RA. RAD denotes denoised RA. RADL is used for LPC spectral-warped RAD. RADC is used for childrenized RD. RADS denotes RAD processed using only the spectral warping part of the proposed method. RADS is used to compare with RADL, since both of them modify only the spectral envelope. RADN denotes RADS along with the proposed pitch shifting, or RADC without the proposed vowel stretching. It is used to analyze the effect of pitch shifting and vowel stretching.

3.1.3. Automatic speech recognition system

The main objective of the proposed data augmentation method is to improve ASR performance for children's speech. We build a Hybrid HMM/TDNN acoustic models using the Kaldi toolkit [21]. The acoustic model training follows the Kaldi's WSJ *s5* recipe with iVectors. For language modelling, we train a Kneser-Ney trigram language model using the SRILM toolkit [22]. The language models also utilize options to prune and limit word length sizes for training. The language models are trained using Icelandic Gigaword corpus 2021 [23].

3.1.4. Child-adult speech classification system

Child-adult speech classification (CASC) task is used to measure how childrenization mimics real children. Childrenized adult utterances are used as instances of children to train a CASC model. If the childrenzied data well mimic real children, the model should be able to correctly classify a decent amount of real children utterances, meanwhile confusing little adult utterances. Unweighted accuracy (UA) is used as the evaluation metric for CASC.

A two-dimensional convolutional neural network (CNN) is used to classify if the speaker of an utterance is an adult or a child. The CNN architecture used in this work is similar to VGG-M [24], widely used for image classification and speech-related applications [25]. The input features of the neural network are 300 frames long 80-dimensional mel-spectrogram images computed with a frame length of 30 ms and a hop length of 10 ms using Librosa [26]. The CNN model maps each mel-spectrogram into a 400 vector speaker embedding.

Given a test utterance, the model predicts the probability of the test utterance belonging to a child and an adult speaker. We select the score value with the highest probability and assign the test utterance to either a child or adult speaker. Since the input of the classification model is 3 seconds audio clip, we keep only utterances that are larger than 3 seconds in the test sets. Thus, the number of test files is reduced from 1714 to 1174.

3.2. Experimental Results

We present the Word Error Rates (WER) from our ASR experiments and Unweighted Accuracy (UA) on our child-adult

Table 1: *Word Error Rates (WER) in ASR and Unweighted Accuracy (UA) of the Child Adult Speaker Classification (CASC) on the test splits of Malromur, a real adult speech corpus, and Samromur, a real children speech corpus. For details refer to Section 3.1.1 and Section 3.1.2.*

	ASR WER		**CASC UA**
Training	Adult	Child	
RA (*109-hrs real adult*)	9.0	51.3	-
+ RAL (*LPC*)	8.8	46.5	17.9
+ RAC (*childrenized*)	9.0	40.7	52.7
+ RADL (*denoised + LPC*)	**7.9**	44.6	33.6
+ RADS (*denoised + s. w.*)	8.3	43.6	57.4
+ RADN (*denoised + s. w. + p. s.*)	8.4	38.5	63.0
+ RADC (*denoised + childrenized*)	8.8	**36.3**	**68.8**
+RC$_5$ (*real children 5.4-hrs subset*)	8.9	28.7	89.4
+ RC$_5$+RADC	8.7	26.5	91.8
RC (*127-hrs real children*)	22.7	16.6	-
RA + RC (*236 hrs*)	9.1	17.6	99.6
Wav2Vec2.0 (*1001 hrs*) [27]	5.7	9.4	

Table 2: *Confusion matrix on the child-adult speaker classification task for models developed using the reference (RA+RADL) and the proposed (RA+RADC) data augmentation methods.*

Training	Child RADL	Adult RA	Child RADC	Adult RA
Testing	Hyp Child	Hyp Adult	Hyp Child	Hyp Adult
Ref Child	99	1075	529	645
Ref Adult	4134	5905	736	9303

speaker classification in Table 1.

The last three rows of the table display the oracle numbers achievable when domain-matched real children (RC) data is used for training our ASR or the Icelandic Wav2Vec2.0 [27] system, which was trained on 1001 hours of Icelandic speech including children speech. Obtaining such an amount of labelled children speech is difficult for most languages, and hence we focus on the scenario where only real adult data is available. As a naïve baseline, we train our model on adult speech (RA) and test on children speech.

Evaluating ASR on the children's test set, we observe 9.3% relative improvement over the baseline when augmenting with LPC-based generated data (RA+RAL), which has been earlier used to augment data for such children speech tasks [4]. In comparison to the same test set, we observe larger gains (20.6% relatively) when augmenting with our childrenization scheme (RA+RAC), showing that the latter scheme is better.

Next, we also experiment denoising the input adult data before childrenization (refer Fig. 1). The proposed speech enhancement, or denoising, substantially improves the childrenization-based scheme and it also slightly improves the LPC-based scheme.

Then we experimented with the three aspects of modifications used in the proposed childrenization method on denoised adult data. Performing only spectral warping (RA+RADS) enabled ASR improvement in both children and adults. Including pitch shift (RA+RADN) further improved children ASR to a large extent, meanwhile resulting in a slight performance drop in adult ASR. Full childrenization, including also vowel stretching (RA+RADC), further improved children ASR but slightly harmed adult ASR.

In terms of adult ASR performance, the above models perform similar to or better than the naïve baseline. LPC-based spectral warping with denoising (RA+RADL) is the most effective data augmentation method on adults (approximately 12 % relative improvement). This suggests that LPC-based spectral warping with denoising can be used as a generic data augmentation method to improve the training data variability.

In CASC experiments, using the proposed method (RADC) to produce children instances resulted in an Unweighted Accuracy (UA) of 68.8%. The detailed confusion matrix in Table 2 shows 45.1% children utterances being correctly recognized meanwhile confusing only 7.3% of adult utterances. This suggests that the proposed childrenization method can mimic real children to some extent. In comparison, using the LPC-based method (RADL) to produce children instances resulted in a UA of 33.6%, with only 8% children utterances being correctly recognized meanwhile confusing 41.2% of adult utterances. We made an observation on a few samples, a more aggressive warping factor made LPC-warped utterances perceptually much more childlike. However, a more aggressive warping factor led to less gain in children ASR performance in [4, Table 3].

On both ASR and CASC, we see that the performance gap to best-achievable numbers can be further reduced just by adding a small amount of real children data (RA+RC$_5$+RADC vs RA+RC$_5$). This suggests that a little real data with synthetic data can be complementary to reduce the gap to the matched-data-trained models. Notably, both RC$_5$ and the test real children data are from the Samrómur. A small amount of out-of-domain children data might be less effective.

4. Conclusions

We proposed a data augmentation scheme for children ASR, where we aimed at converting adult speech to childlike speech by leveraging speech enhancement and voice conversion techniques. Using our proposed method, we converted labelled adult speech to be childlike and augmented ASR training data for scenarios where no labelled real children data or very little labelled real children data is available.

Training on publicly-available Icelandic adult data and evaluating on publicly-available Icelandic children speech, we improved children WER by 29% in comparison to the adult-only model by including synthetic child data in training. This is close to half of the 65% improvement when including real child data in training. In our experiments, we noted that the proposed speech enhancement step was an essential part of the process, improving the childrenized adult speech quality for ASR. The proposed method also substantially outperformed the reference LPC-based method that is aimed at improving children ASR. In addition, we evaluated child-adult speaker classification to see the resemblance of childrenized adult data to real children. In absence of real children data for training, our method taking childrenized adult speech as instances of the children class outperformed different baselines including our reference model based on the LPC-based adult speech. Overall, our proposed method best reduces the gap to the oracle models in comparison to other data augmentation methods exemplifying the quality of the generated childlike speech for children speech tasks.

In this paper, we investigated a voice conversion method that required no in-domain training data. The resulting speech was sometimes not as clean as with deep-learning-based voice conversion methods. In future, we will investigate parameter-rich methods to produce childlike speech in out-of-domain tasks.

5. References

[1] S. B. Lovato and A. M. Piper, "Young children and voice search: What we know from human-computer interaction research," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00008

[2] A. Potamianos and S. S. Narayanan, "Robust recognition of children's speech," *IEEE Transaction on Audio Speech and Language Processing*, vol. 11, no. 6, pp. 603–616, 2003. [Online]. Available: https://doi.org/10.1109/TSA.2003.818026

[3] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, and C. Y. Espy-Wilson, "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 455–459. [Online]. Available: https://doi.org/10.21437/Interspeech.2016-1399

[4] H. K. Kathania, V. Kadyan, S. R. Kadiri, and M. Kurimo, "Data augmentation using spectral warping for low resource children asr," *Journal of Signal Processing Systems*, p. 7, 2022. [Online]. Available: http://urn.fi/URN:NBN:fi:aalto-202211236653

[5] H. K. Kathania, A. Kumar, and M. Kurimo, "Vowel non-vowel based spectral warping and time scale modification for improvement in children's ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 6983–6987. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9414116

[6] V. P. Singh, H. B. Sailor, S. Bhattacharya, and A. Pandey, "Spectral modification based data augmentation for improving end-to-end ASR for children's speech," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 3213–3217. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-11343

[7] 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, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4382–4386. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-1112

[8] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, "Data augmentation using CycleGAN for end-to-end children ASR," in *29th European Signal Processing Conference, EUSIPCO 2021, Dublin, Ireland, August 23-27, 2021*. IEEE, 2021, pp. 511–515. [Online]. Available: https://doi.org/10.23919/EUSIPCO54536.2021.9616228

[9] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[10] 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. E99.D, no. 7, pp. 1877–1884, 2016.

[11] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639314000697

[12] ——, "D4c, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639316300413

[13] ——, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 2321–2325. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0068.html

[14] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of Acoustical Society of America*, vol. 105(3), pp. 1455–1468, 1999.

[15] A. Nicolson and K. Paliwal, "On training targets for deep learning approaches to clean speech magnitude spectrum estimation." *The Journal of the Acoustical Society of America*, vol. 149(5).

[16] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, 2020.

[17] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1803.01271*, 2018.

[18] C. D. Hernandez Mena, D. E. Mollberg, M. Borský, and J. Gunason, "Samrómur children: An Icelandic speech corpus," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 995–1002. [Online]. Available: https://aclanthology.org/2022.lrec-1.105

[19] C. Mena, M. Borsky, D. E. Mollberg, S. F. Gumundsson, S. Hedström, R. Pálsson, Ó. H. Jónsson, S. orsteinsdóttir, J. V. Gumundsdóttir, E. H. Magnúsdóttir, R. órhallsdóttir, and J. Gudnason, "Samromur children 21.09," 2021, CLARIN-IS. [Online]. Available: http://hdl.handle.net/20.500.12537/185

[20] S. Steingrímsson, J. Gunason, S. Helgadóttir, and E. Rögnvaldsson, "Málrómur: A manually verified corpus of recorded Icelandic speech," in *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, May 2017, pp. 237–240. [Online]. Available: https://aclanthology.org/W17-0229

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[22] A. Stolcke, "Srilm - an extensible language modeling toolkit." in *INTERSPEECH*, J. H. L. Hansen and B. L. Pellom, Eds. ISCA, 2002. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2002.html#Stolcke02

[23] S. Barkarson, S. Steingrímsson, H. Hafsteinsdóttir, . D. Andrésdóttir, I. G. Eiríksdóttir, B. Magnússon, and F. Ingimundarson, "The Icelandic gigaword corpus (IGC) 2021," 2021, CLARIN-IS. [Online]. Available: http://hdl.handle.net/20.500.12537/192

[24] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[25] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[26] B. McFee, M. McVicar, C. Raffel, D. Liang, O. Nieto, J. Moore, D. Ellis, D. Repetto, P. Viktorin, J. F. Santos *et al.*, "Librosa: v0.4.0," *Zenodo 2015*, 2015.

[27] C. D. Hernandez Mena, "Acoustic model in icelandic: wav2vec2-large-xlsr-53-icelandic-ep10-1000h." 2022. [Online]. Available: https://huggingface.co/carlosdanielhernandezmena/wav2vec2-large-xlsr-53-icelandic-ep10-1000h