Journal of Signal Processing Systems (2022) 94:1507–1513
https://doi.org/10.1007/s11265-022-01820-0

Data Augmentation Using Spectral Warping for Low Resource Children ASR

Hemant Kumar Kathania[1,2] · Viredner Kadyan[3] · Sudarsana Reddy Kadiri[1] · Mikko Kurimo[1]

Received: 22 June 2022 / Revised: 4 October 2022 / Accepted: 9 October 2022 / Published online: 8 November 2022

Abstract

In low resource children automatic speech recognition (ASR) the performance is degraded due to limited acoustic and speaker variability available in small datasets. In this paper, we propose a spectral warping based data augmentation method to capture more acoustic and speaker variability. This is carried out by warping the linear prediction (LP) spectra computed from speech data. The warped LP spectra computed in a frame-based manner are used with the corresponding LP residuals to synthesize speech to capture more variability. The proposed augmentation method is shown to improve the ASR system performance over the baseline system. We have compared the proposed method with four well-known data augmentation methods: pitch scaling, speaking rate, SpecAug and vocal tract length perturbation (VTLP), and found that the proposed method performs the best. Further, we have combined the proposed method with these existing data augmentation methods to improve the ASR system performance even more. The combined system consisting of the original data, VTLP, SpecAug and the proposed spectral warping method gave the best performance by a relative word error rate reduction of 32.13% and 10.51% over the baseline system for Punjabi children and TLT-school corpus, respectively. The proposed spectral warping method is publicly available at https://github.com/kathania/Spectral-Warping.

Keywords Children speech recognition · Spectral warping · Prosody modification · VTLP · SpecAug · TDNN

1 Introduction

The growing trend of internet use and service digitization demands speech based applications for entertainment, games and education for children as well as adults. In the development of voice user interfaces, the lack of data resources poses a serious concern for automatic speech recognition (ASR) technologies. This low resource issue typically concerns transcribed speech data and linguistic knowledge, because a successful ASR engine typically requires thousands of hours of training data to sufficiently cover the large variability in acoustics, vocabulary and speaking style [1].

For developing ASR systems, many Indian languages face the challenge of data scarcity. While various approaches are employed to handle data scarcity issues [2], many of these approaches fail because of the small number of speakers in the training data. To better address the data scarcity, data augmentation can be employed to multiply the original data using, for example, generative adversarial networks (GAN) and text-to-speech (TTS) [3–6], or through external or internal augmentation on various acoustic methods [7, 31]. Such augmentations are only successful either through real time resource collection or by production of data which is similar to natural speech [8]. Other common approaches for data augmentation are prosody modification and speed/volume perturbation [2, 9–11]. Apart from these, in [12] spectral augmentation was used directly on

✉ Hemant Kumar Kathania
hemant.ece@nitsikkim.ac.in

✉ Sudarsana Reddy Kadiri
sudarsana.kadiri@aalto.fi

Viredner Kadyan
vkadyan@ddn.upes.ac.in

Mikko Kurimo
mikko.kurimo@aalto.fi

[1] Department of Signal Processing and Acoustics, Aalto University, Espoo 02150, Finland

[2] Department of Electronics and Communication Engineering, National Institute of Technology Sikkim, Ravangla 737139, India

[3] Speech and Language Research Centre, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India

Mel frequency cepstral coefficient (MFCC) features and filter bank features to improve ASR performance.

For building children ASR, it is particularly difficult to cover the larger acoustic and linguistic variation due to the limited availability of training data [13–15]. Previous studies have shown that large variations in formant frequencies, fundamental frequency and spectral variability degrades the recognition performance. In order to address these, the focus has been on age dependent acoustic modeling and vocal tract length normalization (VTLN) [13, 16–19]. Few studies have been presented on adding spectral variation either by augmentation or through complete replacement of the original data [7, 12, 18, 20]. Recently, spectral warping has been presented in mismatched data condition where the ASR models are trained with adults and tested with children speech data [7]. In a custom way, the method was shown to improve ASR performance in mismatched data conditions, but its effectiveness has not yet been investigated on low resource languages. Now, in this study we investigate the effectiveness of the spectral warping based data augmentation in low resource children ASR.

In this paper, we collected a corpus of children's speech data for the Indian Punjabi language. Because of the limited amount of training data, the baseline TDNN-based ASR results were poor. To address the data scarcity and also to capture more acoustic and speaker variability from speech production point of view (among children of different ages), we proposed a spectral warping based data augmentation method. We compared our proposed method with four well-known existing methods: pitch scaling [9, 10], speaking rate [9, 10], SpecAug [12] and VTLP [21] based data augmentations and found that the proposed method performs the best.

The main highlights of this study are as follows:

1. Proposed a simple and efficient spectral warping based data augmentation for improving the performance of children ASR in low resource conditions.
2. Systematic investigation involving two speech databases: Punjabi language data (collected in this study in India) and TLT-school data (non-native children corpus [22]).
3. Systematic comparison between the proposed spectral warping based data augmentation and four well-known existing data augmentation methods (pitch scaling, speaking rate modification, SpecAug and VTLP).
4. Investigation of complementary information among the data augmentations by various combinations of the proposed and the existing data augmentation methods.

2 Databases Used

Two low resource children speech databases were used: Punjabi language and TLT-school non-native English [22].

2.1 Punjabi Language Children Speech Database

This corpus has been collected in this study from native speakers (Punjabi language) of Punjab state of India. The corpus was built by reading aloud Punjab School Education board books. The corpus consists of speech data of school students in the age group of 7-14 years. The database was collected from 79 speakers (35 male and 44 female). The data was divided into train and test parts using the 80:20 ratio. The database details are given in Table 1. The collected speech varied with respect to different prosody parameters such as speaking rate, and pitch. Speakers up to 10 years spoke long sentences by concatenating or leaving a short gap between two successive sentences, while speakers above 10 years read aloud fluently. The entire corpus was collected under a controlled clean acoustic environment through mobile device. The sampling frequency of the collected data is 16 kHz.

2.2 TLT-school Non-native Children Speech Database

To check the robustness of the proposed data augmentation, we also experimented with the TLT-school non-native children speech database [22]. The database contains 49.20 hours of speech from 3112 speakers for training, and 2.20 hours of speech from 84 speakers for testing. The age range of the TLT-school speakers varies from 9 to 16 years. The details of the database are given in Table 1.

3 Baseline ASR System

We used a Kaldi toolkit recipe (https://github.com/kaldi-asr/kaldi) to train the baseline ASR system [23]. MFCC features were computed with a Hamming window of 20 ms through a frame shift of 10 ms. This front-end employed a Mel-filter bank of 40 channels which finally produced 13-dimension static feature vectors. These features were further time spliced with ± 4 frames in left and right of the current frame which results into a 117-dimension feature vector for each frame.

Table 1 Details of the Punjabi children speech database and TLT-school non-native children speech database.

Data		Training	Testing
Punjabi	No. of speakers	63	16
	Speaker age	7-14 years	7-14 years
	Duration (hrs.)	12.20	2.50
TLT-School	No. of speakers	3112	84
	Speaker age	9-16 years	9-16 years
	Duration (hrs.)	40.20	2.20

This larger dimension of features was later reduced by linear discriminant analysis (LDA) along with Maximum likelihood linear regression (MLLR) to generate 40-dimension vectors. Further, these vectors were decorrelated using Cepstral mean and variance normalization (CMVN) and speaker variability issues were reduced using Feature space maximum likelihood linear regression (fMLLR). The acoustic models were built on these vectors using hidden Markov models (HMMs) with deep neural network (DNN) output observation probabilities.

Initially, experiments were performed using the DNN-HMM based hybrid acoustic model. This approach leaves the sequence modelling for HMM and focuses on the good representation power of DNN in the short acoustic contexts. Consequently, a limited number of hidden layers were varied in the original system to find the best architecture for the remaining experiments. The system gave the best results with *tanh* function by employing 3 layers and each hidden layer consists of 512 units. The initial learning rate was kept as small as 0.005 and the final learning rate was 0.0005. Apart from the DNN, a time delay neural network (TDNN) [24] acoustic model was also investigated which tries to model sequences also with the feed-forward architecture. The first layer of the TDNN architecture processes only a narrow context of the input and the next layers concatenate the outputs of the hidden activations to cover a wider time window. The target is to efficiently learn wider temporal relationships. We also used i-vectors [25] for TDNN. To decode the test set, a domain specific language model (LM) was built. This LM was trained on all transcripts of the train utterances. Baseline word error rate (WERs) for DNN and TDNN-based systems are given in Table 2 for two databases. WER is the most popular metric used for ASR system evaluation. It measures the percentage of incorrect words. It is defined as:

$$WER = \frac{S + D + I}{N}. \tag{1}$$

Where S = Total number of substitutions, D = Total number of deletions, and I = Total number of Insertions. From the results in table, it can be observed that the TDNN system clearly outperforms the DNN system for both the databases. Hence, further experiments in this study are carried out with the TDNN-based acoustic model.

Table 2 Baseline WERs obtained on DNN and TDNN-based ASR system for Punjabi and TLT-School databases.

System Type	WER (%)	
	Punjabi	TLT-School
DNN	12.73	25.43
TDNN	9.18	20.26

4 Spectral Warping Based Data Augmentation

To increase the spectral variability, the spectral structure of the children's speech data is modified using the spectral warping approach. This is carried out using the spectrum obtained with linear prediction (LP) method. It is expected that modifying the features derived from the warped spectrum of the speech signal provides useful spectral variability to improve the ASR performance.

The warping of the LP spectrum (denoted by $X_\beta(f)$) is carried out from the original LP spectrum (denoted by $X(f)$) computed from children's speech using a warping function $V_\beta(f)$. Here β is the warping factor.

$$X_\beta(f) = X(V_\beta(f)).$$

According to the traditional LP method, an estimate of the current speech sample $x(n)$ can be obtained as a linear combination of past K speech samples, which is given by:

$$\hat{x}(n) = \sum_{k=1}^{K} a_k x(n-k). \tag{2}$$

The Z-transform of Eq. (2) is obtained as:

$$\hat{X}(z) = \left(\sum_{k=1}^{K} a_k z^{-k}\right) X(z). \tag{3}$$

Here X(z) and $\hat{X}(z)$ denote the Z-transforms of the prediction signal $\hat{x}(n)$ and the speech signal $x(n)$, respectively. The LP coefficients are denoted by a_k and z^{-k} denote the k-unit delay filters.

The warping to the LP spectrum is applied by replacing the unit delay filters with a first order all-pass filter A(z). The first order filter is given by [26–28]:

$$A(z) = \frac{z^{-1} - \beta}{1 - \beta z^{-1}}. \tag{4}$$

The range of the warping factor is: $-1 < \beta < 1$. With the warping function A(z) on the LP coefficients a_k, the spectral structure of the LP spectra can be shifted systematically. The positive values of β shift the entire spectrum towards lower frequencies, i.e., the left side, and on the other hand the negative β values shift the entire spectrum towards higher frequencies, i.e., the right side. This is illustrated with the LP spectrum for a segment of voiced speech in Fig. 1, where the red curve shows the original LP spectrum, the blue curve shows the warped LP spectrum for $\beta = 0.1$, and the green curve shows the warped LP spectrum for $\beta = -0.1$.

The spectral warped speech signal is derived using the warped LP coefficients, a_k, with the residual $(x(n) - \hat{x}(n))$ using a classical LP synthesizer [29]. The synthesized speech signal with the proposed approach is referred to as the *spectral warped* (SW) speech signal in the present study. This spectral warped speech signal is used for data augmentation.

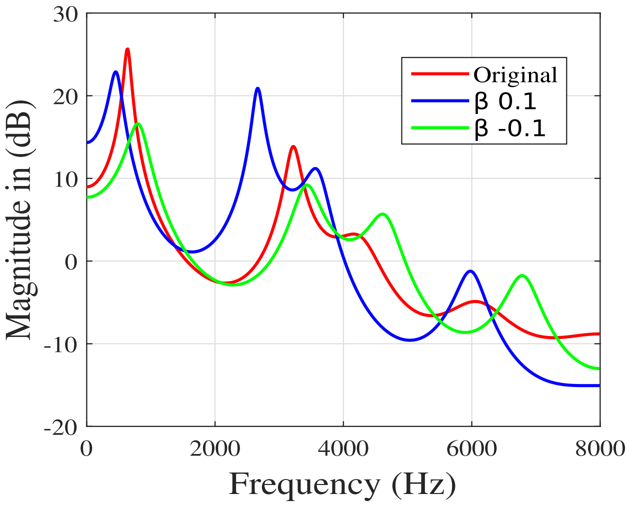

Figure 1 An illustration of LP spectrum (red curve) for a segment of voiced speech. The warped LP spectrum for $\beta = 0.1$, and $\beta = -0.1$ are shown in blue and green curves.

To overcome the issue of data scarcity that affects the ASR performance, we used the proposed data augmentation method to create spectral warped (SW) data and merged that with the original data to create the augmented training data as shown in Fig. 2. Spectral warping was performed by tweaking its tunable parameter (β) from −0.1 to 0.1 with a step size of 0.05. We used the augmented data for system training using different spectral warping parameter β on train set and the results are given in Table 3 for both the databases. From the table, it can be observed that the lowest WER was obtained for $\beta = -0.05$, and it gave a relative improvement of 9% for Punjabi and 3% for TLT-school over the TDNN baseline system. In the remaining experiments, $\beta = -0.05$ is used. Cross-validation was not used as test data is so small.

5 Comparison with Well-known Data Augmentation Methods

To test the effectiveness of the proposed spectral warping (SW) based data augmentation over the existing data augmentation methods, this section gives the performance comparison of ASR systems on Punjabi and TLT-School databases. Four well-known and popular data augmentation methods are considered for this purpose. They are: pitch scaling [9, 10], speaking rate modification [9, 10], SpecAug [12] and VTLP [21], which have been shown to improve the performance of ASR systems. Prosodic parameters such as pitch and speaking rate were varied to leverage prosodic variation in the data. In this study, pitch scaling (PS) and speaking rate (SR) are modified with Time Scale Modification (TSM) using the Real-Time Iterative Spectrogram Inversion with Look-Ahead (RTISI-LA) algorithm [9, 10, 30]. This algorithm constructs a high-quality time-domain speech signal from its short-time magnitude spectrum. Both the parameters (PS and SR) are tunable and the best value was searched by varying them from 0.65–1.45 in PS and 0.65–1.85 in SR, with a step size of 0.1. The best value for pitch was found to be 0.85 and for speaking rate was 1.15, and these values are utilized for further investigation of the ASR performance. In SpecAug data augmentation, spectrogram is modified by removing time and frequency information randomly [12]. In VTLP, warping factor value of 0.90 and 1.10 for each utterance were varied to leverage vocal tract length variation in the data [21]. We then augment the modified data with each of the method to the original data to train an ASR system. The results (WERs and relative improvements) obtained for proposed data augmentation (SW) and existing methods such as, PS, SR, SpecAug and VTLP are

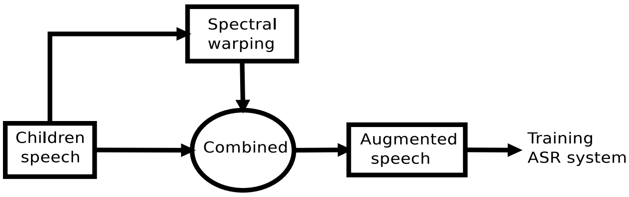

Figure 2 Block diagram of the proposed spectral warping augmentation process.

Table 3 WERs (and relative improvement over baseline) obtained on spectral warping (SW) based data augmentation for Punjabi and TLT-School databases

System Type	WER (%)		Relative Imp. (%)	
	Punjabi	TLT-School	Punjabi	TLT-School
Original (O)	9.18	20.26	–	–
O+SW(+0.1)	8.83	20.12	3.81	0.69
O+SW(+0.05)	8.67	19.98	5.55	1.38
O+SW(-0.05)	**8.34**	**19.73**	9.15	2.61
O+SW(-0.1)	8.39	19.86	8.60	1.97

Bold numbers indicates the best WER

Table 4 Performance comparison (WERs and relative improvement over baseline) of proposed (SW) data augmentation with well-known data augmentation methods for Punjabi and TLT-school databases

System Type	WER (%)		Relative Imp. (%)	
	Punjabi	TLT-school	Punjabi	TLT-school
Original (O)	9.18	20.26	–	–
+ Pitch (PS)	8.98	19.98	2.17	1.36
+ Speaking rate (SR)	9.06	20.05	1.30	1.03
+ SpecAug	8.47	19.84	7.73	2.07
+ VTLP	8.53	19.89	7.08	1.82
+ Spectral warping (SW)	**8.34**	**19.73**	9.15	2.61

Bold numbers indicates the best WER

Table 5 WERs (and relative improvement over baseline) obtained on data augmentation combinations: O+SW+PS, O+SW+SR, O+SW+SpecAug, and O+SW+VTLP for Punjabi and TLT-School databases.

System Type	WER (%)		Relative Imp. (%)	
	Punjabi	TLT-school	Punjabi	TLT-school
Baseline	9.18	20.26	–	–
O+SW+PS	6.63	19.06	27.77	5.92
O+SW+SR	8.32	19.37	9.36	4.63
O+SW+SpecAug	6.54	18.87	28.78	6.86
O+SW+VTLP	6.96	18.74	24.18	7.50

given in Table 4. From the table, it can be observed that all the data augmentation methods improved the system performance over the baseline. Among the existing methods, SpecAug and VTLP gave a larger improvement than PS and SR. Overall, the proposed SW method gave a larger relative improvement than any of the four existing data augmentation methods.

6 Combining Spectral Warping and Existing Data Augmentation Methods

To further capture more acoustic and speaker variability, and to enhance the system performance for low resource data, we combined the proposed spectral warping based data augmentation with the PS, SR, SpecAug, and VTLP based data augmentations. These experiments show the effectiveness and complementary nature of the proposed SW method with the existing data augmentation methods. First, experiments are carried out with the combination of the proposed (SW) method and each one of the existing data augmentation method (PS, SR, SpecAug or VTLP) along with original (O) data, to train ASR systems. The experiments are: O+SW+PS, O+SW+SR, O+SW+SpecAug, and O+SW+VTLP. The results (WERs and relative improvements) of the experiments are given in Table 5. From the table, it can be observed that all the combinations improved the system performance over the baseline, indicating the complementary nature of the proposed (SW) method with each of the existing data augmentation method. Among the combinations, O+SW+SR gave a smaller improvement compared to the other three combinations. Hence in later combination experiments, SR is not considered.

As the results in Table 5 indicated that proposed SW method has complementary information with existing methods, further experiments are carried out by combining multiple data augmentations which include the proposed SW method along with original data to train ASR systems. The combined experiments are: O+SW+PS+SpecAug, O+SW+PS+VTLP, O+SW+SpecAug+VTLP, and O+SW+PS+SpecAug+VTLP. The results (WERs and relative improvements) of the combined experiments are given in Table 6. From the table, it can be observed that all the combinations improved the system performance over the baseline, indicating the complementary nature of the proposed (SW) method with the combinations of the existing data augmentation methods. Among the combinations, O+SW+SpecAug+VTLP gave a better performance compared to the other three combinations. This best system gave a relative improvement of 32.13% and 10.51% compared to baseline system for the Punjabi and TLT school databases, respectively. We have also conducted a statistical significance test for the best model (O+SW+SpecAug+VTLP) and noticed that the signed pair comparison found significant difference between the baseline system and best combined system at level $p < 0.01$.

Table 6 WERs (and relative improvement over baseline) obtained on combination of O+SW+PS+SpecAug, O+SW+PS+VTLP, O+SW+SpecAug+VTLP, and O+SW+PS+SpecAug+VTLP for Punjabi and TLT-School databases

System Type	WER (%)		Relative Imp. (%)	
	Punjabi	TLT-school	Punjabi	TLT-school
Baseline	9.18	20.26	–	–
O+SW+PS+SpecAug	6.46	18.53	29.62	8.53
O+SW+PS+VTLP	6.31	18.46	31.26	8.88
O+SW+SpecAug+VTLP	**6.23**	**18.13**	32.13	10.51
O+SW+PS+SpecAug+VTLP	6.39	18.32	30.39	9.57

Bold numbers indicates the best WER

7 Conclusion

In this paper, a spectral warping (SW) based data augmentation method was proposed to improve the ASR performance in low resource children speech databases. The effectiveness of the proposed SW method was investigated on two databases (Punjabi and TLT-School), and it was shown to improve the ASR performance (reduction in WER) as compared to the four existing data augmentation methods (PS, SR, SpecAug and VTLP). Further, complementary information among the data augmentations was found from the experiments on combination of the proposed (SW) and the existing data augmentation methods. The best system (O+SW+SpecAug+VTLP) gave a relative improvement of 32.13% and 10.51% compared to the baseline system for the Punjabi and TLT school databases, respectively. Note that the proposed spectral warping based data augmentation is a simple and efficient signal processing method and it is not dependent on database. Hence it can be used for other language databases as well to improve the performance of ASR systems.

Acknowledgements This work was supported by the Academy of Finland (grants 329267, 330139).

Author Contributions Conceptualization, Hemant Kumar Kathania, Viredner Kadyan and Sudarsana Reddy Kadiri.; methodology, Hemant Kumar Kathania, Viredner Kadyan and Sudarsana Reddy Kadiri.; software, Hemant Kumar Kathania, Viredner Kadyan and Sudarsana Reddy Kadiri; validation, Hemant Kumar Kathania, Viredner Kadyan, Sudarsana Reddy Kadiri and Mikko Kurimo; formal analysis, Hemant Kumar Kathania, Viredner Kadyan and Sudarsana Reddy Kadiri; investigation, Hemant Kumar Kathania, Viredner Kadyan and Sudarsana Reddy Kadiri; resources, Hemant Kumar Kathania and Viredner Kadyan; data curation, Hemant Kumar Kathania and Viredner Kadyan; writing—original draft preparation, Hemant Kumar Kathania, Viredner Kadyan and Sudarsana Reddy Kadiri; writing—review and editing, Hemant Kumar Kathania, Viredner Kadyan and Sudarsana Reddy Kadiri and Mikko Kurimo.; visualization, Hemant Kumar Kathania, Viredner Kadyan and Sudarsana Reddy Kadir; supervision, Mikko Kurimo; project administration, Mikko Kurimo; funding acquisition, Mikko Kurimo. All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding provided by Aalto University. Academy of Finland (grants 329267, 330139).

Data Availability Publicly available datasets were analyzed in this study. These data can be found here: http://www.thespeechark.com/pf-star-page.html and https://catalog.ldc.upenn.edu/LDC95S24. Indian Punjabi data can be obtained by requesting the second author (Virender Kadyan).

Code Availability https://github.com/kathania/Spectral-Warping.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Yes, all authors have read and agreed to the for publication.

Conflicts of Interest Authors do not have any Conflict of interest/Competing interests.

References

1. Evermann, G., Chan, H. Y., Gales, M. J., Hain, T., Liu, X., Mrva, D., Wang, L., & Woodland, P. C. (2004). Development of the 2003 CU-HTK conversational telephone speech transcription system. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (vol. 1, p. 249). IEEE.
2. Kanda, N., Takeda, R., & Obuchi, Y. (2013). Elastic spectral distortion for low resource speech recognition with deep neural networks. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 309–314). IEEE.
3. Hu, H., Tan, T., & Qian, Y. (2018). Generative adversarial networks based data augmentation for noise robust speech recognition. In *Proceedings - ICASSP* (pp. 5044–5048).
4. Qian, Y., Hu, H., & Tan, T. (2019). Data augmentation using generative adversarial networks for robust speech recognition. *Speech Communication, 114,* 1–9.
5. Gales, M. J., Ragni, A., AlDamarki, H., & Gautier, C. (2009). Support vector machines for noise robust ASR. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 205–210). IEEE.
6. Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., & Wu, Z. (2019). Speech recognition with augmented synthesized speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 996–1002). IEEE.
7. Kathania, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2020). Study of formant modification for children ASR. In *Proceedings - ICASSP* (pp. 7429–7433). IEEE.
8. Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.-N., Mohamed, A., & Dupoux, E. (2021). Speech resynthesis from discrete disentangled self-supervised representations. Preprint retrieved from https://arxiv.org/abs/2104.00355
9. Shahnawazuddin, S., Adiga, N., Kathania, H. K., & Sai, B. T. (2020). Creating speaker independent ASR system through prosody modification based data augmentation. *Pattern Recognition Letters, 131,* 213–218.
10. Kathania, H., Singh, M., Grósz, T., & Kurimo, M. (2020). Data augmentation using prosody and false starts to recognize non-native children's speech. In *Proceedings of the Annual Conference of the International Speech Communication Association,*

INTERSPEECH (pp. 260–264). https://doi.org/10.21437/Interspeech.2020-2199

11. Kathania, H. K., Shahnawazuddin, S., Ahmad, W., Adiga, N., Jana, S. K., & Samaddar, A. B. (2018). Improving children's speech recognition through time scale modification based speaking rate adaptation. In *2018 International Conference on Signal Processing and Communications (SPCOM)*.
12. Park, D. S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. In G. Kubin, & Z. Kacic (Eds.) *INTERSPEECH 2019*, (pp. 2613–2617). ISCA, Graz, Austria.
13. Lee, S., Potamianos, A., & Narayanan, S. S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105*(3), 1455–1468.
14. Narayanan, S., & Potamianos, A. (2002). Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing, 10*(2), 65–78.
15. Kumar, V., Kumar, A., & Shahnawazuddin, S. (2022). Creating robust children's ASR system in zero-resource condition through out-of-domain data augmentation. *Circuits, Systems, and Signal Processing, 41,* 2205–2220.
16. Potaminaos, A., & Narayanan, S. (2003). Robust Recognition of Children Speech. *IEEE Transactions on Speech and Audio Processing, 11*(6), 603–616.
17. Serizel, R., & Giuliani, D. (2014). Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In *2014 IEEE Spoken Language Technology Workshop (SLT)* (pp. 135–140).
18. Jaitly, N., & Hinton, G. E. (2013). Vocal tract length perturbation (VTLP) improves speech recognition. In *ICML Workshop on Deep Learning for Audio, Speech and Language*.
19. Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(9), 1469–1477. https://doi.org/10.1109/TASLP.2015.2438544
20. Shahnawazuddin, S., Kumar, A., Kumar, V., Kumar, S., & Ahmad, W. (2022). Robust children's speech recognition in zero resource condition. *Applied Acoustics, 185*, 108382.
21. Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015* (pp. 3586–3589). ISCA. http://www.isca-speech.org/archive/interspeech_2015/i15_3586.html
22. Gretter, R., Matassoni, M., Bannò, S., & Falavigna, D. (2020). TLT-school: A corpus of non native children speech. CoRR *abs/2001.08051*. https://arxiv.org/abs/2001.08051
23. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech recognition toolkit. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).*
24. Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In B. Yegnanarayana (Ed.) *INTERSPEECH 2018* (pp. 3743–3747). ISCA, Hyderabad, India.
25. Saon, G., Soltau, H., Nahamoo, D., & Picheny, M. (2013). Speaker adaptation of neural network acoustic models using I-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013* (pp. 55–59). IEEE.
26. Strube, H. W. (1980). Linear prediction on a warped frequency scale. *The Journal of the Acoustical Society of America, 68*(4), 1071–1076.
27. Laine, U.K., Karjalainen, M., & Altosaar, T. (1994). Warped linear prediction (WLP) in speech and audio processing. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing* (vol. 3, p. 349). IEEE.
28. Smith, J. O., & Abel, J. S. (1999). Bark and erb bilinear transforms. *IEEE Transactions on speech and Audio Processing, 7*(6), 697–708.
29. Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE, 63*(4), 561–580.
30. Zhu, X., Beauregard, G. T., & Wyse, L. L. (2007). Real-time signal estimation from modified short-time fourier transform magnitude spectra. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(5), 1645–1653.
31. Kathania, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2022). A formant modification method for improved ASR of children's speech. *Speech Communication, 136*, 98–106.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.