




Spectral warping based data augmentation for low resource children's speaker verification

Hemant Kumar Kathania^{1,2} · Virender Kadyan³ · Sudarsana Reddy Kadiri¹  · Mikko Kurimo¹

Received: 5 November 2022 / Revised: 17 August 2023 / Accepted: 22 September 2023 /

Published online: 3 November 2023

© The Author(s) 2023

Abstract

In this paper, we present our effort to develop an automatic speaker verification (ASV) system for low resources children's data. For the children's speakers, very limited amount of speech data is available in majority of the languages for training the ASV system. Developing an ASV system under low resource conditions is a very challenging problem. To develop the robust baseline system, we merged out of domain adults' data with children's data to train the ASV system and tested with children's speech. This kind of system leads to acoustic mismatches between training and testing data. To overcome this issue, we have proposed spectral warping based data augmentation. We modified adult speech data using spectral warping method (to simulate like children's speech) and added it to the training data to overcome data scarcity and mismatch between adults' and children's speech. The proposed data augmentation gives 20.46% and 52.52% relative improvement (in equal error rate) for Indian Punjabi and British English speech databases, respectively. We compared our proposed method with well known data augmentation methods: SpecAugment, speed perturbation (SP) and vocal tract length perturbation (VTLP), and found that the proposed method performed best. The proposed spectral warping method is publicly available at <https://github.com/kathania/Speaker-Verification-spectral-warping>.

✉ Sudarsana Reddy Kadiri
sudarsana.kadiri@aalto.fi

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

Virender Kadyan
vkadyan@ddn.upes.ac.in

Mikko Kurimo
mikko.kurimo@aalto.fi

¹ Department of Information and Communications Engineering, 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

Keywords Speaker verification · Children’s speech · Spectral warping · Low resource languages · Speed perturbation · Vocal tract length perturbation

1 Introduction

Automatic speaker verification (ASV) of children’s speakers has many potential applications in child security and protection, games, education and entertainment. In these applications, the performance of the deployed ASV system is affected by various factors. It is well known that the acoustic and linguistic characteristics of children’s speakers are gradually improving/changing as increasing the age. Another important issue affecting ASV of children’s speakers is the limited amount of publicly available children speech data [1, 2], where as for adults’ speech, there are databases including more than 1000 hours of training data to train ASV systems [3–5]. These acoustic and linguistics changes of children speech together with the lack of training data make ASV more challenging.

In initial studies [6, 7], it was found that young speakers aged three to thirteen had age-dependent variations in formants and fundamental frequency measurements. Children’s voices have higher fundamental and formant frequencies, as well as greater spectral diversity, as compared to adults’ voices. In-Domain and out-of-Domain data augmentations techniques were explored in [8]. In-Domain data augmentations were carried out using speaking rate and pitch modification and out-Domain data augmentations were carried out using combination of adults and children to improve a child ASV system in a limited children data scenario and found the proposed approach give reduction in equal error rate (EER) [8]. Authors also explored the voice conversion (VC) approach to modify the adults’ speech to resemble the children’s speech using cycle-consistent generative adversarial network (GAN). Experimental comparisons were carried out using both x-vector and i-vector-based speaker modeling in the context of children’s ASV. Further, age-group wise analysis was carried out to see the effect of data augmentation on ASV performance with variations in age of the speakers.

In [9], vocal tract information was used for children’s speaker verification and it was shown to improve the ASV system performance. Explanation for degraded recognizer scores through acoustic changes resulting from voice disguise is presented in [10]. In [11], synthesis based data augmentation method was used to expand the training set with text-controlled synthesized speech for low resource data, and was shown a reduction in EER. The voice model compression technique is described in [12] to improve the system performance for low resource scenario. In [13], fast binary features were explored for speaker recognition and they were found to improve the ASV system performance. In [14], speaker verification system based on a shared neural embedding space for adults and children was presented and the system achieved promising results with adults and children.

The X-vector based speaker recognition was investigated for short utterances and it was observed that the system significantly improved the performance [15]. Data augmentation can also be beneficial for imbalanced classes of data. For instance, in [16], various deep learning techniques are investigated for tackling class-imbalanced data. The survey reveals that dealing with highly unbalanced data presents extra challenges, as most learners tend to exhibit a bias towards the majority class and, in severe instances, might entirely disregard the minority class. In another work [17], the author explores whether larger data is always superior and demonstrates its impact on the proper and improper utilization of big data in machine learning approaches.

In this paper, we collected a corpus of children’s speech data for the Indian Punjabi language. 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 two well-known existing data augmentation methods: namely, speed perturbation (SP) [8, 18, 19] and vocal tract length perturbation (VTLP) [20], and found that the proposed method performs the best.

The main highlights of this study are as follows:

- Proposed a simple and efficient spectral warping based data augmentation for improving the performance of children ASV in low resource conditions.
- Systematic investigation involving two speech databases: Punjabi language data (collected in this study in India for both adults’ and children’s), and British English WSJCAM0 (adults’) [21] PFSTAR (children’s) [22] databases .
- Systematic comparison between the proposed spectral warping based data augmentation and two well-known existing data augmentation methods (speed perturbation and VTLP).

2 Speech databases

To check the robustness of the proposed data augmentation for children speaker verification, experiments were carried out with two language databases: (1) Indian Punjabi and (2) British English.

2.1 Indian Punjabi corpus

Punjabi corpus has been collected from native speakers (Punjabi language) of Punjab state of India for both children’s and adults’ speakers [23]. The corpus was built through generation of read speech using utterances taken from Punjab School Education board books. The data was collected for a total of 66 children’s speakers (denoted as P-Children) and 47 adult speakers (denoted as P-Adult). The age range of children lies in the 7-14 years and adults’ speakers in the 18-28 years. In total, children’s data consists of 11.46 hrs and adults’ speakers of 17.09 hrs. The children’s speech corpus was divided into train and test parts using the 80:20 ratio. The entire corpus was collected under a controlled clean acoustic environment through mobile devices. The sampling frequency of the collected data is 16 kHz. The database details like age group, duration, etc., are given in Table 1.

2.2 British English corpus

Experiments were also carried out with British English speech corpus. For adults’ speakers, WSJCAM0 database [21] is considered and PFSTAR [22] is used for children’s speakers. In total WSJCAM0 consists of 15.5 hrs of data from 92 speakers with an age range of greater than 18-44 years with sampling frequency of 16 kHz, and PFSTAR consists of 13.18 hrs of data from 134 speakers with age range 4-14 years with sampling frequency of 16 kHz. The children speech corpus was divided nearly into 80:20 ratio for train and test parts. Details about the WSJCAM0 and PFSTAR databases are given in Table 1.

Table 1 Details of two databases (Indian Punjabi speech corpus and British English corpus) used for developing ASV systems

Purpose	Language					
	Indian Punjabi			British English		
	Adult	Children		Adult (WSJCAM0)	Children (PFSTAR)	
	Training	Training	Testing	Training	Training	Testing
No. of speakers	47	53	13	92	107	27
Speaker age	18-28 years	7-14 years	7-14 years	18-45years	4-14 years	4-14 years
Duration (hrs.)	17.09 hrs	10.19 hrs	1.27 hrs	15.50 hrs	10.06 hrs	1.47 hrs

3 Baseline speaker verification system and results

Kaldi recipe (<https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb>) was used to develop a baseline automatic speaker verification (ASV) system [24]. Baseline ASV system is built by employing MFCC front end approach. To extract the MFCC feature vector a short time frames information has been overlapped using hamming window. Non-voice regions are removed from speech signals through a voice activity detector (VAD). A hamming window of 20 ms with a half percentage of overlapping factor on short-term frames are employed. A 30 channel log Mel-filterbanks was used to derive the mel-spectrum and a 30 dimensional MFCCs are extracted after applying DCT on warped mel-spectrum. A time-delay neural network (TDNN) [25] was used to derive fixed-length vectors (referred to as x-vectors) representing the speaker-specific information from acoustic MFCC features (along with 5 frame context). TDNN consists of 7 hidden layers that process the short-time speech MFCC features with a softmax as output layer. Rectified linear units (ReLU) non-linearities are employed in the hidden layers. After processing the features through hidden layers, the output is fed to statistics pooling layer, and mean and standard deviation over time are computed. The output of pooling layer is propagated to other hidden layers. An affine transform followed by non-linear activation is applied to the output of the pooling layer to derive the speaker embedding x-vector. Finally scoring is performed using x-vectors using probabilistic linear discriminant analysis (PLDA) model and cosine distance. The system evaluations are performed using Equal Error Rate [26].

A baseline system was built for matched condition (i.e., the system trained with children's speech and tested with children's speech) and mismatched condition (i.e., the system trained with adults' speech and tested with children's speech) and EERs are reported in Table 4 for both the databases, namely, Indian Punjabi corpus and British English corpus as discussed in Section 2. From the table, it can be observed that the performance of matched case is superior as compared to mismatch case for Punjabi database. Whereas in case of English database matched case EER is slightly higher than mismatched case, this is due to higher amount of data in the training for mismatched case. To make a robust ASV system, we merged adults' speech and children's speech for training the ASV system, and tested with the children's speech for both the databases, and the results are reported in Table 4. From the results in table, it can be noted that the merged system gives reduction in EERs with a relative improvement of 18.85% and 57.15% over matched case for Indian Punjabi and British English corpus, respectively. This robust ASV system is considered as a baseline system for further study. Even though the performance of the system is improving there exists acoustic mismatches between training and testing data. To overcome this issue, we have proposed to

modify the adults' speech like children's speech using spectral warping method to use as an augmentation, which is discussed in next section (Section 4).

4 Spectral Warping based data augmentation

To increase the amount of data and to overcome data scarcity issue, we proposed a spectral modification method to augment the data. Spectral modification method enhances the spectral variability, i.e., the spectral structure of the adults' speech data is modified using the spectral warping approach. This is carried out using the warping of spectrum of linear prediction (LP) method [27, 28]. It is hypothesized that extraction of features from the warped spectrum of the speech signal provides useful spectral variability to improve the performance of speaker verification system.

The warping of the LP spectrum denoted by $X_\beta(f)$, is carried out from the LP spectrum denoted by $X(f)$ of adults' speech using a warping function $V_\beta(f)$, where β is the warping factor.

$$X_\beta(f) = X(V_\beta(f)). \quad (1)$$

According to the conventional LP method, an estimate of the present speech sample ($x(n)$) can be derived as a linear combination of past K speech samples. This is given by:

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

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

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

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

Warping to the LP spectrum is done by replacing the unit delay filter with a all-pass filter $A(z)$, which is a first-order filter. This is given by [29–31]:

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

The value of β (warping factor) is in the range of $-1 < \beta < 1$. With the variation of warping function $A(z)$ on the LP coefficients (a'_k 's), the spectral structure of the LP spectra can be modified or shifted systematically. A positive value of β shifts the entire spectrum towards lower frequencies, i.e., the left side. On the other hand, the negative value of β shifts the entire spectrum towards higher frequencies, i.e., the right side. This phenomenon 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/modified speech signal is reconstructed using the warped LP coefficients, a'_k 's, with the residual ($x(n) - \hat{x}(n)$) using a traditional LP synthesizer [32]. The synthesized speech signal is referred to as the *spectral warped* speech signal in this study. This spectral warped speech signal is used to augment the training data of the speaker verification system. The code is made publicly available at <https://github.com/kathania/Speaker-Verification-spectral-warping>.

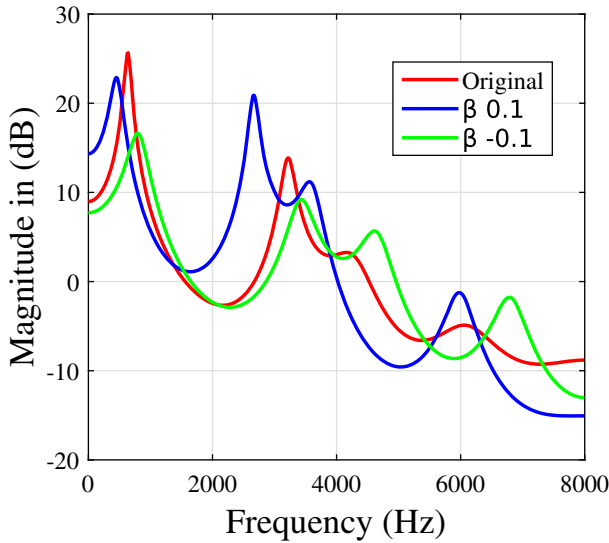


Fig. 1 An illustration of LP spectrum (red curve) for a segment of voiced speech. The warped LP spectrum for $\beta = 0.1$, and $\beta = -0.1$ are shown in blue and green curves

To overcome the issue of data scarcity which affects the performance of ASV system, the proposed data augmentation is used to create spectrally warped (SW) speech and is augmented with the original data to create the training data. Block diagram of proposed spectral warping based data augmentation for automatic speaker verification (ASV) system is given in Fig. 2. Spectral warping was performed by tweaking its tunable parameter (β) from -0.25 to 0.20 with a step size of 0.05. The data is augmented for system training with different spectral warping parameters (β).

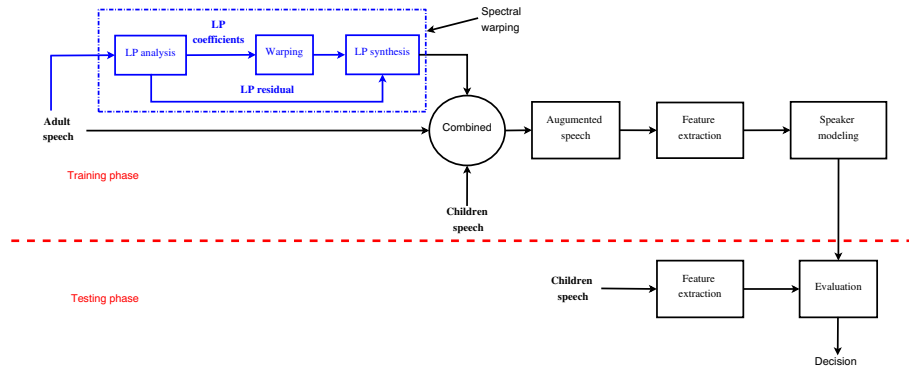


Fig. 2 Block diagram of proposed spectral warping based data augmentation for automatic speaker verification (ASV) system

5 Results and discussion

This section first describes the results of proposed spectral warping (SW) based data augmentation and then the effectiveness of the proposed data augmentation over the existing data augmentation methods is described for both Indian Punjabi and British English databases. Two well-known and popular data augmentation methods are considered for comparison purpose, they are: speed perturbation (SP) [18, 19], and vocal tract length perturbation (VTLP) [20], which have been shown to improve the performance of ASV systems. Finally, the complementary information among the data augmentation methods is described.

5.1 Results of spectral warping based data augmentation

It is evident from the results in Section 3 that the baseline ASV system is effective for more training data. Although merged system improve the ASV system performance, there exists a mismatch between training and testing data, and scarcity of training data for children speech. We modified adults' speech like children's speech using spectral warping method discussed in Section 4. Spectral warping was performed by tweaking its tunable parameter (β) from -0.25 to 0.20 with a step size of 0.05 to modify the adults' speech to like children's speech for overcoming the mismatch condition between training and testing database. Experiments are conducted by training the modified adults' speech with β varying from -0.25 to 0.20 and testing with children's speech for both the Indian Punjabi and British English databases. The results are shown in Fig. 3. From the figure, it can be noted that the proposed method

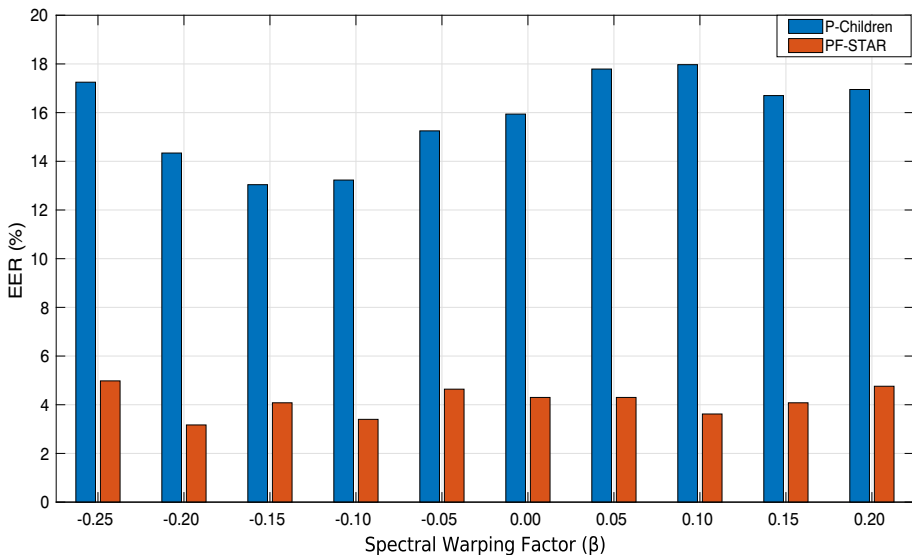


Fig. 3 Demonstration of proposed spectral warping (SW) to overcome the mismatch condition during training and testing. Results obtained with training data of modified adults' speech (like children's speech) using spectral warping by varying β from -0.25 to 0.20 with a step size of 0.05, and testing with children speech. Here P-Children refers to Indian Punjabi children's speech and PF-STAR refers to British English children speech

Table 2 Performance of ASV system (EER) for matched condition (i.e., the system trained with children’s speech and tested with children’s speech), mismatched condition (i.e., the system trained with adults’ speech and tested with children’s speech) and merged (i.e., the system trained with combination of adults’ and children’s speech, and tested with children’s speech) condition for Indian Punjabi and British English databases

System Type	Training	Testing	EER	R.I (%)
Punjabi Data				
Matched	P-Children	P-Children	8.01	-
Mismatched	P-Adult	P-Children	15.94	-
Merged	P-Children+P-Adult	P-Children	6.50	18.85
English Data				
Matched	PFSTAR	PFSTAR	5.55	-
Mismatched	WSJCAM0	PFSTAR	4.3	-
Merged	PFSTAR + WSJCAM0	P-Children	2.38	57.15

R.I indicates the relative improvement (in %) over matched case

Bold indicates best performance

improve the system performance compared to no spectral modification (see results in Table 2 for mismatched condition) applied for both the databases. The best performance is achieved for the β value of -0.10 and -0.15 for Punjabi database, and for British English database the best performance is achieved for the β value of -0.10 and -0.20.

To overcome data scarcity and as well as mismatch between training and testing, spectral warping modified adults’ data of two best warping factor β is augmented with baseline system (children’s + adults’) as a training data (denoted as, baseline+spectral warping (SW)). We trained the ASV model with augmented (baseline+spectral warping (SW) data with two best warping factor β) for both databases with testing children’s data, and the results are reported in Table 3. From the table, it can be observed that spectral warping based data augmentation improve the ASV system performance for both the Indian Punjabi and British English databases. The augmentation methods gives relative improvement of 20.46% and 52.52% as compared to baseline system for Indian Punjabi and British English databases, respectively.

Table 3 Performance of ASV system (EER) for baseline (i.e., the system trained with adults’ and children’s speech, and tested with children’s speech), and proposed spectral warping (SW) (i.e., training data of baseline+SW and testing with children’s speech) for Indian Punjabi and British English databases

Training	Testing	EER (%)	R.I.(%)
Punjabi Data			
Children’s + Adults’ (baseline)	P-Children	6.50	-
Baseline	P-Children	5.17	20.46
+ Spectral Warping (SW)			
English Data			
PFSTAR + WSJCAM0 (baseline)	PFSTAR	2.38	-
Baseline	PFSTAR	1.13	52.52
+ Spectral Warping (SW)			

R.I indicates the relative improvement (in %) over baseline. This demonstrates the effectiveness of proposed spectral warping (SW) to overcome the mismatch condition and data scarcity

Bold indicates best performance

Table 4 Performance of ASV system (EER) on merged system (i.e., the system trained with combination of adults' and children's speech) and proposed spectral warping (SW) (i.e., training with baseline+SW and testing with children's speech of different age-groups) for Indian Punjabi and British English databases

Training	Testing	EER	R.I (%)
Punjabi Data			
P-Children+P-Adult	P-Children	6.50	
P-Children+P-Adult	P-Children(4-6 years)	No-data	-
P-Children+P-Adult	P-Children(7-9 years)	6.94	-
P-Children+P-Adult + Spectral Warping (SW)	P-Children(7-9 years)	5.23	24.63
P-Children+P-Adult	P-Children(10-14 years)	6.12	-
P-Children+P-Adult + Spectral Warping (SW)	P-Children(10-14 years)	5.09	16.83
English Data			
PFSTAR + WSJCAM0	PFSTAR	2.38	-
PFSTAR + WSJCAM0	PFSTAR (4-6 years)	7.32	-
PFSTAR + WSJCAM0 + Spectral Warping (SW)	PFSTAR (4-6 years)	3.36	54.09
PFSTAR + WSJCAM0	PFSTAR (7-9 years)	2.72	-
PFSTAR + WSJCAM0 + Spectral Warping (SW)	PFSTAR (7-9 years)	1.20	55.88
PFSTAR + WSJCAM0	PFSTAR (10-14 years)	1.93	-
PFSTAR + WSJCAM0 + Spectral Warping (SW)	PFSTAR (10-14 years)	1.02	47.15

R.I indicates the relative improvement (in %)

Bold indicates best performance

Further, we have studied the performance of the proposed method for different age-groups of both the databases, Indian Punjabi and British English. For each database, we have divided the test data into age groups of 4-6, 7-9, and 10-14 years. The age-wise analysis is reported in Table 4, and it is found that for each age group, the proposed method gave a reduction in EER. The best improvement was found for the age group of 7-9 years in both databases.

5.2 Comparison with existing data augmentation methods

To analyze the effectiveness of proposed augmentation method with existing data augmentation methods, three well known augmentation methods are explored. They are: speed perturbation (SP) [19, 34, 35], vocal tract length perturbation (VTLP) [20], and SpecAugment [33] which have been shown to improve the performance of ASV and automatic speech recognition systems. For SP, VTLP, and SpecAugment, we have used Kaldi recipe [35]. The speed of speech signal is modified with factors of 0.90 and 1.10. In VTLP, warping factor values of 0.90 and 1.10 are used to leverage vocal tract length variation in the data [20]. On other side, in SpecAugment, the spectrogram is modified so that its time and frequency information is removed randomly. The modified data with each of the method is augmented to the baseline system to train an ASV system. The results (EERs and relative improvements (R.I)) obtained for proposed data augmentation (SW) and existing methods such as, SP, VTLP and SpecAugment are given in Table 5 for both the Indian Punjabi and British English databases. From the table, it can be clearly seen that all the data augmentation methods improved the system performance over the baseline ASV system. Among the existing methods, SpecAugment

Table 5 Performance of ASV systems (in EER) for proposed spectral warping (SW) and existing data augmentations (SpecAugment, speed perturbation (SP) and vocal tract length perturbation (VTLP)) for Indian Punjabi and British English databases

Method	Training	Testing	EER(%)	R.I.(%)
Punjabi Data				
Baseline	P-Children + P-Adult	P-Children	6.5	-
Spectral Warping (SW)	P-Children + P-Adult + SW	P-Children	5.17	20.46
SpecAugment [33]	P-Children + P-Adult + SpecAug	P-Children	5.29	18.61
Speed Perturbation (SP) [19, 34, 35]	P-Children + P-Adult + SP	P-Children	5.38	17.23
VTLP [20]	P-Children + P-Adult + VLTP	P-Children	5.61	13.69
English Data				
Baseline	PFSTAR + WSJCAM0	PFSTAR	2.38	-
Spectral Warping (SW)	PFSTAR + WSJCAM0 + SW	PFSTAR	1.13	52.52
SpecAugment [33]	PFSTAR + WSJCAM0 + SpecAug	PFSTAR	1.31	44.95
Speed Perturbation (SP) [19, 34, 35]	PFSTAR + WSJCAM0 + SP	PFSTAR	1.52	36.13
VTLP [20]	PFSTAR + WSJCAM0 + VLTP	PFSTAR	1.67	29.83

R.I indicates the relative improvement (in %) over the baseline

Bold indicates best performance

gave better performance in comparison to VTLP and SP. Overall, the proposed SW method gave a larger relative improvement of 20.46% and 52.52% for Indian Punjabi and British English databases, respectively, than any of the three existing data augmentation methods.

6 Conclusion

Developing ASV systems for children's speakers is a challenging task because of limited data availability of children's speech. In this work, we have developed a robust baseline system with merging children's speech with adult's speech data to train ASV systems and testing with children's speech. This type of system leads to mismatches between training and testing, and reduces the performance. A spectral warping based data augmentation method studied to overcome this issue. Using this method, we convert adult speech towards children's speech and added to baseline system to overcome the mismatch conditions between training and testing. Two speech databases namely Indian Punjabi and British English speech corpus are used to show the effectiveness of our proposed method. It was shown that the proposed method improved the system performance for both the databases compared to the baseline system. We have also compared our proposed method with existing methods: SpecAugment, speed perturbation (SP) and vocal tract length perturbation (VTLP), and found the proposed method performed best. This study used the single warping factor for all the age groups. It will be interesting to investigate with the optimal warping factor with age-wise. In future, the effectiveness of proposed method can be explored for additional languages apart from Punjabi and English that are used in this study. Further studies can be made to see the effectiveness of proposed method on noisy speech.

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 Kadiri; 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).

Declarations

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

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Claus F, Gamboa-Rosales H, Petrick R, Hain H-U, Hoffmann R (2013) A survey about databases of children's speech. In: Proc. INTERSPEECH, pp 2410–2414
2. Fainberg J, Bell P, Lincoln M, Renals S (2016) Improving children's speech recognition through out-of-domain data augmentation. In: Proc. INTERSPEECH 2016:1598–1602
3. Nagrani A, Chung JS, Zisserman A (2017) Voxceleb: a large-scale speaker identification dataset. In: Proc. INTERSPEECH
4. Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: An ASR corpus based on public domain audio books. In: Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP):pp 5206–5210
5. Battenberg E, Chen J, Child R, Coates A, Gaur Y, Li Y, Liu H, Satheesh S, Seetapun D, Sriram A, Zhu Z (2017) Exploring neural transducers for end-to-end speech recognition. CoRR [arXiv:1707.07413](https://arxiv.org/abs/1707.07413)
6. Eguchi S, Hirsh IJ (1969) Development of speech sounds in children. *Acta oto-laryngologica. Supplementum* 257:1–51
7. Kent RD (1976) Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *JHSR* 9:421–447
8. Shahnawazuddin S, Ahmad W, Adiga N, Kumar A (2020) In-domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario. In: Proc. ICASSP, pp 7554–7558
9. Safavi S, Najafian M, Hanani A, Russell M, Jancovic P, Carey M (2012) Speaker recognition for children's speech. In: Proc. INTERSPEECH, vol 3

10. González Hautamäki R, Hautamäki V, Kinnunen T (2019) On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. *J Acoust Soc Am* 146(1):693–704
11. Du C, Han B, Wang S, Qian Y, Yu K (2021) Synaug: Synthesis-based data augmentation for text-dependent speaker verification. In: Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5844–5848
12. Tydilat B, Navratil J, Pelecanos JW, Ramaswamy GN (2007) Text-independent speaker verification in embedded environments. In: Proc. IEEE international conference on acoustics, speech and signal processing - ICASSP, vol 4, pp 293–296
13. Laptik R, Sledevi T (2017) Fast binary features for speaker recognition in embedded systems. In: Proc. Open conference of electrical, electronic and information sciences (eStream) pp 1–4
14. Kaseva T, Kathania HK, Rouhe A, Kurimo M (2021) Speaker verification experiments for adults and children using a shared embedding spaces. In: Proc NoDaLiDa 2021, pp 86–93
15. Kanagasundaram A, Sridharan S, Sriram G, Prachi S, Fookes C (2019) A study of x-vector based speaker recognition on short utterances
16. Johnson KTMJM (2019) Survey on deep learning with class imbalance. *J Big Data*. springer vol 6
17. Rocchetti DGCLM (2019) Is bigger always better a controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. *J Big Data*. Springer, vol 6
18. Shahnawazuddin S, Adiga N, Kathania HK, Sai BT (2020) Creating speaker independent ASR system through prosody modification based data augmentation. *Pattern Recogn Lett* 131:213–218
19. Kathania H, Singh M, Grósz T, Kurimo M (2020) Data augmentation using prosody and false starts to recognize non-native children’s speech. In: Proc INTERSPEECH 2020, pp 260–264
20. Ko T, Peddinti V, Povey D, Khudanpur S (2015) Audio augmentation for speech recognition. In: INTER-SPEECH 2015, 16th annual conference of the international speech communication association. Dresden, Germany, September 6–10, pp 3586–3589
21. Robinson T, Franssen J, Pye D, Foote J, Renals S (1995) WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In: Proc. ICASSP, vol 1, pp 81–84
22. Batliner A, Blomberg M, D’Arcy S, Elenius D, Giuliani D, Gerosa M, Hacker C, Russell M, Wong M (2005) The PF_STAR children’s speech corpus. In: Proc. INTERSPEECH, pp 2761–2764
23. Dua M, Kadyan V, Banthia N, Bansal A, Agarwal T (2022) Spectral warping and data augmentation for low resource language asr system under mismatched conditions. *Appl Acoust* 190:108643
24. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P et al (2011) The kaldi speech recognition toolkit. In: IEEE workshop on automatic speech recognition and understanding
25. Snyder D, Garcia-Romero D, Povey D, Khudanpur S (2017) Deep neural network embeddings for text-independent speaker verification. In: Proc. Interspeech, pp 999–1003
26. Povey D, Zhang X, Khudanpur S (2014) Parallel training of deep neural networks with natural gradient and parameter averaging. [arXiv:1410.7455](https://arxiv.org/abs/1410.7455)
27. Kathania HK, Kadiri SR, Alku P, Kurimo M (2022) A formant modification method for improved asr of children’s speech. *Speech Comm* 136:98–106
28. Kumar Kathania H, Reddy Kadiri S, Alku P, Kurimo M (2020) Study of formant modification for children asr. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 7429–7433
29. Strube HW (1980) Linear prediction on a warped frequency scale. *J Acoust Soc Am* 68(4):1071–1076
30. Laine UK, Karjalainen M, Altsaar T (1994) Warped linear prediction (wlp) in speech and audio processing. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, vol 3, pp 349
31. Smith JO, Abel JS (1999) Bark and erb bilinear transforms. *IEEE Trans Speech Audio Process* 7(6):697–708
32. Makhoul J (1975) Linear prediction: A tutorial review. *Proc IEEE* 63(4):561–580
33. Park DS, Chan W, Zhang Y, Chiu C-C, Zoph B, Cubuk ED, Le QV (2019) SpecAugment: A simple data augmentation method for automatic speech recognition
34. Shahnawazuddin S, Ahmad W, Adiga N, Kumar A (2021) Children’s speaker verification in low and zero resource conditions. *Digit Signal Process* 116:103115
35. Ko T, Peddinti V, Povey D, Khudanpur S (2015) Audio augmentation for speech recognition. In: Proceedings interspeech, pp 3586–3589