Kathania, Hemant; Kadiri, Sudarsana; Alku, Paavo; Kurimo, Mikko

A Formant Modification Method for Improved ASR of Children's Speech

Published in:
Speech Communication

DOI:
10.1016/j.specom.2021.11.003

Published: 01/01/2022

Document Version
Publisher's PDF, also known as Version of record

Speech Communication 136 (2022) 98–106

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

A formant modification method for improved ASR of children's speech

Hemant Kumar Kathania [a,b,*], Sudarsana Reddy Kadiri [a,**], Paavo Alku [a], Mikko Kurimo [a]

[a] Department of Signal Processing and Acoustics, Aalto University, 02150, Finland
[b] Department of Electronics and Communication Engineering, National Institute of Technology Sikkim, 737139, India

ARTICLE INFO

Keywords:
Children's speech recognition
Formant modification
Noise
DNN
TDNN

ABSTRACT

Differences in acoustic characteristics between children's and adults' speech degrade performance of automatic speech recognition systems when systems trained using adults' speech are used to recognize children's speech. This performance degradation is due to the acoustic mismatch between training and testing. One of the main sources of the acoustic mismatch is the difference in vocal tract resonances (formant frequencies) between adult and child speakers. The present study aims to reduce the mismatch in formant frequencies by modifying formants of children's speech to better correspond to formants of adults' speech. This is carried out by warping the linear prediction (LP) spectrum computed from children's speech. The warped LP spectra computed in a frame-based manner from children's speech are used with the corresponding LP residuals to synthesize speech whose formant structure is closer to that of adults' speech. When used in testing of an ASR system trained using adults' speech, the warping reduces the spectral mismatch in speech between training and testing and improves the system performance in recognition of children's speech. Experiments were conducted using narrowband (8 kHz) and wideband (16 kHz) speech of adult and child speakers from the WSJCAM0 and PF_STAR databases, respectively, and by recognizing children's speech using acoustic models trained with adults' speech. The proposed method gave relative improvements of 24% and 11% for the DNN and TDNN acoustic models, respectively, for narrowband speech. For wideband speech, the technique gave relative improvements of 27% and 13% for the DNN and TDNN acoustic models, respectively. The performance of the proposed method was also compared to two speaker adaptation methods: vocal tract length normalization (VTLN) and speaking rate adaptation (SRA). This comparison showed the best recognition performance for the proposed method. We also combined the proposed method with VTLN and SRA, and found that the combined method gave a further reduction in WER. Moreover, our experiments carried out for noisy speech using various types of additive noise and signal-to-noise ratios showed that the proposed method performs well also for degraded speech.

1. Introduction

Automatic recognition of children's speech has many potential applications in education, games and entertainment. In these applications, the performance of the deployed automatic speech recognition (ASR) system is affected by several factors. It is well known that the acoustic and linguistic characteristics of children's speech are greatly different from those of adults' speech and this degrades ASR of children's speech under mismatch conditions (Potamianos and Narayanan, 2003; Cosi, 2009; Narayanan and Potamianos, 2002; Yeung and Alwan, 2018). Mismatch conditions correspond to training the system with adults' speech and testing it with children's speech. Most ASR systems available publicly work well with adults' speech but in the case of children's speech their performance collapses (Schalkwyk et al., 2010; Shivakumar and Georgiou, 2020; Kennedy et al., 2017). The major reason of this performance degradation is the difference between adult and child vocal tracts (Lee et al., 1999; Narayanan and Potamianos, 2002). Another important issue affecting ASR of children's speech is the limited amount of publicly available speech data recorded from child speakers (Claus et al., 2013; Fainberg et al., 2016). For adults' speech, there are databases including more than 1000 h of training data to train ASR systems (Panayotov et al., 2015; Battenberg et al., 2017). However, available databases for children's speech include only a few hours of speech even for major languages such as English (Panayotov et al., 2015; Claus et al., 2013). Due to the strong effect of the mismatch conditions described above on ASR performance and due to the lack of training data from child speakers, there is need for methods which reduce the acoustic mismatch between adults' and children's speech.

* Corresponding author at: Department of Signal Processing and Acoustics, Aalto University, 02150, Finland.
** Corresponding author.
E-mail addresses: hemant.kathania@aalto.fi, hemant.ece@nitsikkim.ac.in (H.K. Kathania), sudarsana.kadiri@aalto.fi (S.R. Kadiri), paavo.alku@aalto.fi (P. Alku), mikko.kurimo@aalto.fi (M. Kurimo).

https://doi.org/10.1016/j.specom.2021.11.003
Received 6 July 2020; Received in revised form 12 August 2021; Accepted 10 November 2021
Available online 27 December 2021

In the context of children's speech-induced mismatch in ASR, the effect of feature learning using a convolutional neural network-based end-to-end acoustic modeling approach was studied in Dubagunta et al. (2019) and the method was shown to give a reduction in WER. Principal component analysis and heteroscedastic linear discriminant analysis based on a low-rank feature projection was explored in Shahnawazuddin et al. (2018) and an improvement in performance was reported. The effect of the filter bank in ASR of children's speech was studied in Kathania et al. (2019) and it was found that the linear-frequency filter bank performed better compared to the mel and inverse-mel filter banks. The effect of acoustic and linguistic differences between adults' and children's speech has been investigated widely (Shahnawazuddin et al., 2015; Kathania et al., 2014; Shivakumar and Georgiou, 2020; Yadav et al., 2018; Shahnawazuddin et al., 2016; Kathania et al., 2018) and it has been observed that the degradation in performance is mainly due to large differences in both the acoustic and linguistic correlates between speech signals of adults and children. In Kathania et al. (2018), prosodic features like loudness, intensity and voice probability were studied and it was found that combining prosodic features with mel-frequency cepstral coefficients (MFCCs) gave a reduction in WER. Prosodic modifications in pitch and speaking rate were explored in Ahmad et al. (2017), Kathania et al. (2018), Shahnawazuddin et al. (2017) and Kathania et al. (2018) to tackle the effect of the mismatch between training and testing. Data augmentation using prosody (pitch and speaking rate) modification based on modifying glottal closure instants was studied in Shahnawazuddin et al. (2020) and an improvement in WER was reported. In Fainberg et al. (2016), data augmentation using stochastic feature mapping to transform out-of-domain adult speech data was explored to improve the system performance for children's speech.

Changes in formant frequencies as a function of age have been explored in many studies (Lee et al., 1999; Huber et al., 1999; Scukanec et al., 1991; Yildirim et al., 2003). Formant frequencies of children's speech have been found to be higher compared to adults' speech (Lee et al., 1999; Scukanec et al., 1991). This is due to the fact that the length of the vocal tract is inversely proportional to formant frequencies: when the vocal tract length increases, the formant frequencies decrease and vice-versa. The range and magnitude of change in formant frequencies are smaller between older age groups than younger ones.

In this paper, we study the modification of formants to improve recognition of children's speech using an ASR system trained with adults' speech. A linear prediction (LP) -based technique is proposed to warp all-pole spectra of children's speech to have formants similar to adults' speech. The warped LP filter is excited with a LP residual to synthesize a formant-modified speech signal which is used to derive MFCC features in testing the ASR system. A preliminary investigation of the formant modification method was published in Kathania et al. (2020). The present study is a sequel to Kathania et al. (2020) extending this previous work of ours in many respects. The topics studied in the current investigation and their justifications are as follows. First, recognition of children's speech is compared between three acoustic models (GMM, DNN and TDNN). In our preliminary study (Kathania et al., 2020), only GMM and DNN were used. By taking advantage of the more effective TDNN acoustic model (Hinton et al., 2012; Peddinti et al., 2015), the current study aims to further improve the recognition performance of children's speech reported in Kathania et al. (2020). Second, as formants of children's speech are generally higher compared to those in adult speech, some of the higher formants in children's speech (particularly the third formant) might be outside the audio bandwidth of narrowband speech. Since this does not happen for wideband speech of children (or for narrowband adult speech), it was considered justified to study the proposed formant modification method in ASR of children's speech using two signal bandwidths (narrowband, sampled at 8 kHz, and wideband, sampled at 16 kHz). Third, in order to compare the proposed LP-based formant modification method with techniques previously used in ASR in similar children's speech-induced mismatch scenarios, two existing techniques (VTLN Claes et al., 1998; Serizel and Giuliani, 2014; Kathania et al., 2013 and SRA Kathania et al., 2018; Zhu et al., 2007) are used as reference methods. Finally, since children's speech signals often contain background noise, it is justified to conduct ASR experiments not only in clean conditions but also for noise-corrupted children's speech. Therefore, we investigated as the final topic how the proposed method is affected in recognition of narrowband and wideband speech corrupted by various types of additive background noise of different signal-to-noise ratio (SNR) values.

The abbreviations used in the study are described in Table 1.

Table 1
List of abbreviations.

ASR	Automatic speech recognition
DNN	Deep neural network
fMLLR	Feature-space maximum likelihood linear regression
GMM	Gaussian mixture mode
HMM	Hidden Markov model
LDA	Linear discriminant analysis
LM	Language model
LP	Linear prediction
MFCCs	Mel-frequency cepstral coefficients
MLLT	Maximum likelihood linear transform
SAT	Speaker adaptive training
SNR	Signal-to-noise ratio
SRA	Speaking rate adaptation
TDNN	Time delay neural network
VTLN	Vocal tract length normalization
WER	Word error rate

2. Formant modification method

The modification of the formant structure of children's speech is carried out by warping the LP spectrum. The warped LP spectrum (denoted by $S_\alpha(f)$) is obtained by modifying the original LP spectrum (denoted by $S(f)$) computed from children's speech using warping function $w_\alpha(f)$, where α is the warping factor:

$$S_\alpha(f) = S(w_\alpha(f)). \tag{1}$$

In conventional LP analysis, an estimate of the present speech sample $s(n)$ is obtained as a linear combination of the past P samples as follows:

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

By Z-transforming Eq. (2), the following equation is obtained

$$\hat{S}(z) = \left(\sum_{k=1}^{P} a_k z^{-k}\right) S(z), \tag{3}$$

where z^{-1} denotes the unit delay filter, $\hat{S}(z)$ and S(z) denote the Z-transforms of the prediction and speech signal, respectively, and a_k are the LP coefficients. In order to warp the LP spectrum, the unit delay filters are replaced by an all-pass filter D(z). The warping of the frequency scale is carried out using a first order filter (Strube, 1980; Laine et al., 1994) given by

$$D(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}. \tag{4}$$

Here α is the warping factor, whose value is in the range of $-1 < \alpha < 1$. The warped frequency scale matches the psychoacoustic frequency scale with a proper selection of α (Smith and Abel, 1999). By taking advantage of the warping function D(z), the spectral resonances (formants) of the LP spectrum can be shifted systematically. For the positive values of α, the formant frequencies shift to lower frequencies. The warped LP coefficients ($a_k's$) can be used to filter the residual (s(n) $-\hat{s}(n)$) to synthesize the speech signal (Makhoul, 1975). The speech

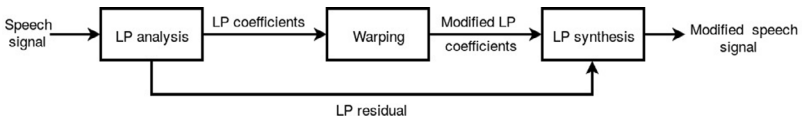

Fig. 1. A block diagram of the LP-based formant modification method.

signal synthesized in this manner is referred to as the *modified* speech signal in the current study. This modified speech signal is used as an input to the ASR system. The steps involved in the proposed method are depicted in the block diagram shown in Fig. 1.

The effectiveness of the proposed formant modification method is illustrated in Fig. 2 (for narrowband speech) and Fig. 3 (for wideband speech) by showing LP spectra computed from vowel utterances spoken by a child and adult speaker (blue and red curves, respectively) together with the modified LP spectra of the child's vowels (green curves) computed by the proposed method. The examples were computed from three vowels representing (according to the standard ARPAbet dictionary) the phonemes /AA/, /OW/ and /AY/, which were taken from reference words "party", "zero", and "my", respectively. The figures demonstrate that the formants of the child speaker are higher compared to those of the adult speaker. Most importantly, the spectra show that the proposed LP-based warping method has moved the formants of the child speaker to be closer to those of the adult speaker. Hence, it is expected that the features derived from the speech signals synthesized using the modified LP spectrum reduce the acoustic mismatch when an ASR system is trained with adults' speech and tested with children's speech. A few demonstration sounds, both with and without the formant modification, have been made available for listening at Github: https://github.com/kathania/Formant-Modification.

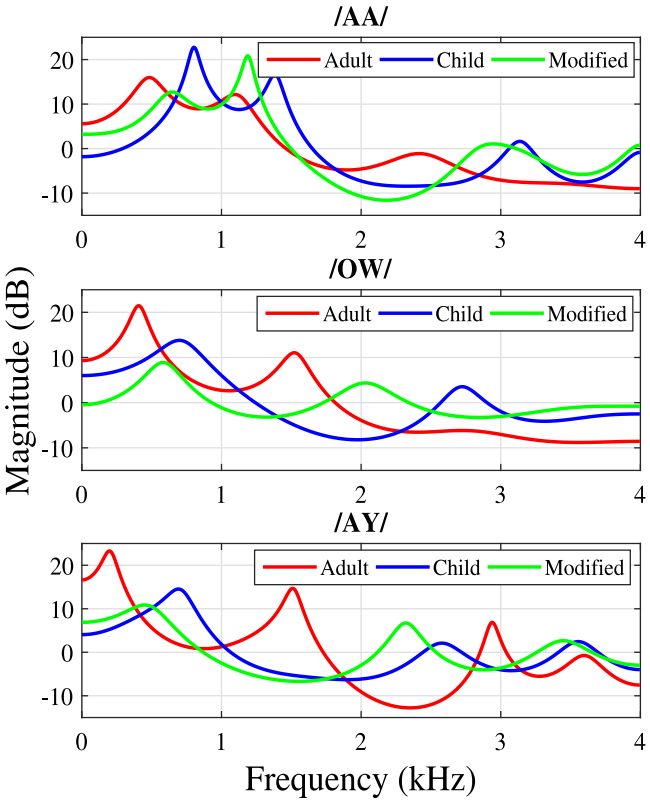

Fig. 2. LP spectra computed from three narrowband vowels (representing the phonemes /AA/, /OW/ and /AY/) showing variation in formant frequencies. The red and blue curves were computed from speech utterances of an adult and child speaker, respectively. The green curves show spectra after applying the formant modification method for the utterances of the child speaker. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Speech data and experimental setup

Two British English speech corpora, WSJCAM0 (Robinson et al., 1995) and PF_STAR (Batliner et al., 2005), were used in the experiments. Both the WSJCAM0 corpus and the PF_STAR corpus contain read speech. The former contains adults' speech and it was used in the current study as a source of training data. The training set of WSJCAM0 contains 15.5 h of speech from 92 speakers (39 females). The PF_STAR corpus contains children's speech. The training set of PF_STAR, which was used for adaptation, consists of 8.3 h of speech from 122 speakers. The test set of PF_STAR includes 1.1 h of speech from 60 speakers (28 females). The age of the child speakers in this corpus varies between 4 and 13 years. The number of speakers, the number of utterances and their total duration in each age group of the test set of PF_STAR are illustrated in Fig. 4. Furthermore, we also built a development set of children's speech for parameter tuning using the PF_STAR training data. The development set of PF_STAR includes 2.1 h of speech from 63 speakers whose age varies between 6 and 14 years.

Performance of most current ASR systems is generally good for wideband speech signals (Russell et al., 2007). However, speech communication in daily life happens often in phone, which involves transmitting the speech signal through narrowband telephone networks. The performance of ASR systems typically degrades for telephone speech due to its narrow bandwidth. The effect of the bandwidth limitation is particularly severe for speech sounds such as fricatives and stops which have prominent spectral components beyond the upper frequency (i.e., 3.4 kHz) of the traditional narrowband. Most importantly, the limitation of the frequency range may be particularly harmful for children's speech because of its higher formants. It is expected that the absence of high-frequency components in narrowband speech of children degrades the ASR performance for child speakers. Therefore, ASR experiments were performed in the current study using both narrowband (sampled at 8 kHz) and wideband speech (sampled at 16 kHz).

The Kaldi toolkit-based recipe was used to perform all the experiments (Povey et al., 2011). Speech data was analyzed in overlapping Hamming-windowed 20 ms frames with a frame-shift of 10 ms to compute MFCC feature vectors. The 40-channel mel-filterbank was used in the MFCC computation. For normalization, cepstral feature-space maximum likelihood linear regression (fMLLR) was used. The fMLLR transformations for the training and test data were generated using the speaker adaptive training (Rath et al., 2013).

Three types of acoustic models were explored in the current study: the Gaussian mixture model (GMM), deep neural network (DNN) and time delay neural network (TDNN). The hidden Markov model (HMM) was used to train the acoustic models. The observation probabilities for the HMM states were generated using the GMM and DNN model (Dahl et al., 2012). Cross-word triphone models consisting of a HMM with eight diagonal covariance Gaussian components per state were used for the GMM–HMM-based ASR system. Furthermore, decision tree-based state tying was performed with the maximum number of tied states (senones) being fixed at 2000. For learning the DNN–HMM-based ASR system, the fMLLR-normalized feature vectors were time-spliced. The number of hidden layers was set to eight with each layer consisting of 1024 hidden nodes. The initial leaning rate for training the DNN–HMM model was set to 0.015 which was reduced to 0.002 after 20

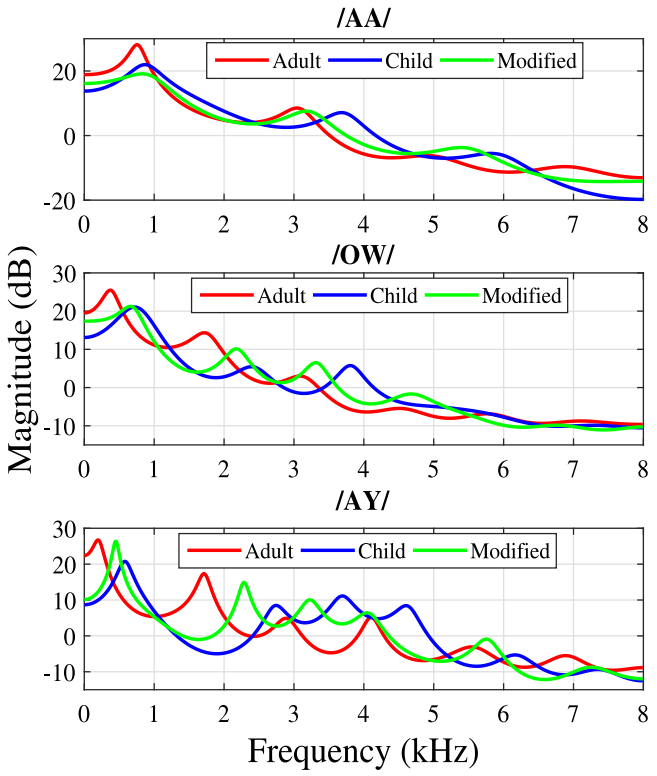

Fig. 3. LP spectra computed from three wideband vowels (representing the phonemes /AA/, /OW/ and /AY/) showing variation in formant frequencies. The red and blue curves were computed from speech utterances of an adult and child speaker, respectively. The green curves show spectra after applying the formant modification method for the utterances of the child speaker. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

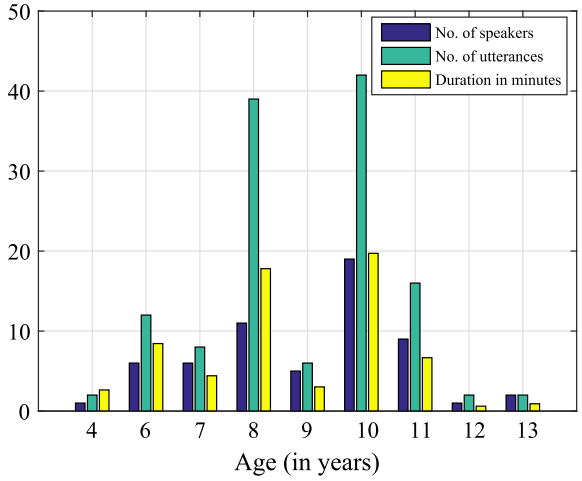

Fig. 4. The number of speakers, the number of utterances and their duration as a function of the speaker age in the PF_STAR test set.

epochs and extra 10 epochs of training were employed. The minibatch size for neural network training was set to 512. The domain-specific language model (LM) was used to decode the children's speech test set. This LM was trained on the transcripts of the speech data of PF_STAR excluding the test set. A portion of the data overlaps between the LM training and testing. A total of 1969 words included in the lexicon with pronunciation variations was employed.

For the TDNN (Povey et al., 2018) acoustic model, the Kaldi toolkit-based setup was used. This setup utilizes MFCC features as input

Table 2
WERs of the baseline ASR system for the GMM and DNN acoustic models.

Acoustic model	Speech bandwidth	WER (%)
GMM	Narrowband	44.65
	Wideband	32.83
DNN	Narrowband	26.23
	Wideband	19.58

to train TDNN-based acoustic models (Povey et al., 2018) on linear discriminant analysis-maximum likelihood linear transform + speaker adaptive training (LDA-MLLT+SAT) based GMM alignment labels. The recipe also performs speaker adaptation of the acoustic model using i-vectors (Saon et al., 2013). Training data was artificially increased 3-fold by time-warping the raw audio. The initial learning rate for training TDNN was set to 0.0005 and finally reduced to 0.00005. The speech recognition results are from two-pass decoding where the first pass utilizes trigram LMs to generate lattices and the second pass rescores the lattices by 4-gram LMs. The perplexity of the PF-STAR test set is 86.5 using these 4-gram LMs.

4. Results

The performance of the baseline ASR system (i.e., the system trained with adults' speech and tested with children's speech without using formant modification) is reported in Table 2 both for the GMM and DNN acoustic models and for narrowband and wideband speech. From the reported WER values it can be concluded that recognition of children's speech with the baseline systems is poor in all the scenarios studied. The general trend of the recognition results is sensible in showing the best result for the DNN model in wideband speech and the worst result for the GMM model in narrowband speech. To study the recognition performance using the proposed formant modification method, WER values are shown in Fig. 5 as a function of the parameter α for the development set of narrowband (upper panel) and wideband (lower panel) speech. It can be seen from the graphs that for narrowband speech the formant modification technique has improved the recognition performance considerably both for the GMM and DNN models and that the improvement is particularly large for the GMM model. For wideband speech, the usage of the formant modification technique has narrowed the gap between the WER values obtained using the GMM model and those obtained using the DNN baseline at $\alpha = 0.1$. In addition, the usage of the formant modification has improved the recognition performance of the DNN model. To be used in the later experiments, we selected the best α value based on the WERs shown in Fig. 5. The data shown in Fig. 5 indicates that the best recognition performance was obtained both for narrowband and wideband speech using $\alpha = 0.1$ and therefore this value was used in all the remaining experiments of the study. Using this α value for the proposed method, further ASR experiments of children's speech were conducted in the following three topics: (1) studying recognition of children's speech with the proposed method in different age group, (2) comparing the proposed method with two adaptation techniques (VTLN and SRA) and (3) studying the effect of noise-corruption on the proposed method. In the following, the experiments of these topics are reported separately in sub-sections.

4.1. Age-wise experiments

In order to study the effect of the age of the child speakers on ASR performance, the test data of children's speech was divided into three different sets based on the speaker's age (4–6 years, 7–9 years, and

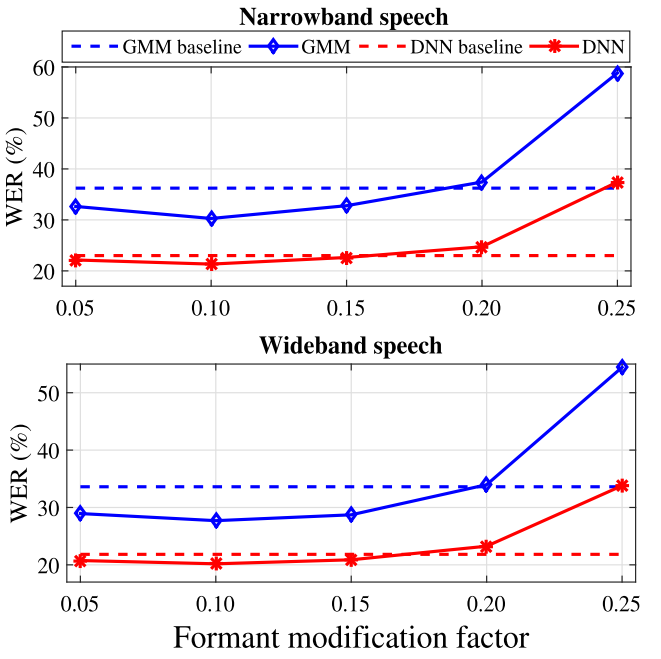

Fig. 5. WERs as a function of the formant modification factor (α) in recognition of children's speech from the development set of PF_STAR in narrowband (upper pane) and wideband (lower pane). The baseline systems are shown by the dotted blue and red lines for the GMM and DNN acoustic models, respectively. The solid blue and red curves show the WERs obtained using the proposed method with the GMM and DNN acoustic models, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
WERs for the age-wise grouped children's speech test sets obtained by the baseline system and the proposed method. The experiments were conducted using the DNN acoustic model. Relative improvement of the proposed method compared to the baseline is also reported.

Age group (in years)	Speech bandwidth	WER (%)		Relative improvement (%)
		Baseline	Proposed	
4–6	Narrowband	91.87	**73.55**	19.94
	Wideband	69.83	**49.33**	29.35
7–9	Narrowband	29.38	**17.26**	41.25
	Wideband	19.12	**10.52**	44.97
10–13	Narrowband	9.54	**9.17**	3.87
	Wideband	11.49	**10.44**	9.13

10–13 years). The experiments were conducted only using the DNN acoustic model due to its better performance compared to the GMM model reported in Table 2. The results of the age-wise experiments are reported in Table 3. It can be observed again that the recognition performance is better for wideband speech compared to narrowband speech. In addition, the results indicate that the proposed method improves the results in all age groups compared to the baseline. It can also be observed that the performance of the system is best for the children in the oldest (10–13 years) age group compared to the two other age groups. The relative improvement of the proposed method compared to the baseline is largest in the middle (7–9 years) age group.

4.2. Comparison to VTLN and SRA

In this section, the proposed method is compared to two existing methods, which have been used in recognition of children's speech: vocal tract length normalization (VTLN) (Claes et al., 1998; Serizel and Giuliani, 2014; Kathania et al., 2013) and speaking rate adaptation (SRA) (Zhu et al., 2007; Kathania et al., 2018). In addition, experiments are conducted using the TDNN acoustic model in addition to the DNN acoustic model which showed the best performance in the experiments reported in Table 2. In the last few years, TDNN has become an effective technique to model the observation densities of the HMM states in comparison to GMM and DNN (Hinton et al., 2012; Peddinti et al., 2015). The WER results of the comparison of the proposed method with VTLN and SRA are reported in Table 4. It can be observed first that the recognition performance improved both for narrowband and wideband speech when the TDNN acoustic model was used instead of the DNN acoustic model. Second, the data shows that both VTLN and SRA improve the recognition performance compared to the baseline corroborating findings reported in previous studies (Claes et al., 1998; Kathania et al., 2013, 2018). Most importantly, the WER values shown in Table 4 indicate that the proposed method gave a clear further improvement in recognition performance compared to VTLN and SRA and that the performance was best for the proposed method with both of the two acoustic models and speech bandwidths.

Additional experiments were conducted by combining the proposed method with the SRA and VTLN techniques. The following three combinations were studied: the combination of the proposed method and VTLN (proposed + VTLN), the combination of the proposed method and SRA (proposed + SRA), and the combination of the proposed method with both VTLN and SRA (proposed + VTLN + SRA). The WER results obtained using these combinations are reported in Table 5 both for narrowband and wideband speech. The data shows that all the combinations yielded better WER values than the proposed method alone, which indicates that VTLN and SRA provide complementary information to the proposed method. The best combination for narrowband speech was proposed + SRA + VTLN, which gave relative improvements of 30% and 22% compared to the baseline with the DNN and TDNN acoustic models, respectively. For wideband speech, the corresponding relative improvements were slightly smaller (14% and 29% for the DNN and TDNN acoustic models, respectively).

As the final step in experimenting with VTLN and SRA, we compared the best of the combined systems (i.e. proposed + VTLN + SRA) with a system which is adapted using children's speech to study the robustness of the combined system. For this purpose, we trained an i-vector extractor (Saon et al., 2013) using wideband children's speech. The results of this experiment are reported in Table 6. From the table, it can be observed that the model adaptation improved the system performance. However, the performance given by the combined system in the corresponding case (i.e., using the TDNN acoustic model for wideband speech, see Table 5) is still better (8.69%) compared to that given by the adapted model (10.64%) reported in Table 6.

In all the previously described experiments, the system was trained using a fairly small database of adult speech, the WSJCAM0 corpus, which contains 15.5 h of speech. We also conducted experiments using a much larger training database, the Librispeech corpus (https://www.openslr.org/12). Librispeech is a widely used speech database in ASR containing 1000 h of speech in US English by 2338 adult speakers (1128 females). The results of these experiments are reported in Table 7. From Table 7, it can be observed that the usage of Librispeech did not improve the system performance compare to the WSJCAM0 database. This is most likely due to the dialect mismatch between the training and testing data (i.e. US English in Librispeech and British English in PF_STAR). However, Table 7 shows that the proposed method improved the system performance compared to the baseline also for Librispeech.

4.3. Experiments with noise-corrupted speech

To study the robustness of the proposed method in noisy conditions, the training and testing data were corrupted with four different

Table 4
WERs obtained by the baseline system, by the VTLN and SRA techniques and by the proposed formant modification method. The experiments were conducted using the DNN and TDNN acoustic models. Relative improvements compared to the baseline are reported for VTLN, SRA and the proposed method.

Acoustic	Speech	WER (%)				Relative improvement (%)		
model	bandwidth	Baseline	VTLN	SRA	Proposed	VTLN	SRA	Proposed
DNN	Narrowband	26.23	23.34	21.98	**19.95**	11.01	16.20	23.94
	Wideband	19.58	15.17	16.68	**14.22**	22.52	14.81	27.37
TDNN	Narrowband	15.83	15.19	14.86	**13.96**	4.04	6.12	11.81
	Wideband	14.16	13.84	13.18	**12.30**	2.25	6.92	13.13

Table 5
WERs obtained by the proposed formant modification method and by combining the proposed method with VTLN and SRA. The experiments were conducted using the DNN and TDNN acoustic models. Relative improvements of the combined systems compared to the baseline are reported.

Acoustic	Speech	WER (%)				Relative improvement (%)		
model	bandwidth	Proposed	Proposed + VTLN	Proposed + SRA	Proposed + VTLN + SRA	Proposed + VTLN	Proposed + SRA	Proposed + VTLN + SRA
DNN	Narrowband	19.95	17.31	15.54	**13.23**	13.29	22.10	30.87
	Wideband	14.22	13.57	13.26	**12.29**	4.57	6.75	13.57
TDNN	Narrowband	13.96	13.23	12.64	**10.87**	5.22	9.45	22.13
	Wideband	12.30	11.54	9.22	**8.69**	6.17	25.04	29.34

Table 6
WERs obtained by the baseline system and by the adapted system based on i-vectors trained using children's speech. The experiments were conducted using the TDNN acoustic model for wideband speech.

WER (%)		Relative
Baseline	Adapted	improvement (%)
14.16	10.64	24.85

Table 7
WERs obtained using the TDNN acoustic model trained on wideband adults' speech from the Librispeech and WSJCAM0 databases, and tested using wideband children's speech from the PF_STAR database. WERs are shown for the baseline and the proposed method for both training databases.

Training	WER (%)		Relative
database	Baseline	Proposed	improvement (%)
Librispeech	23.93	22.33	6.68
WSJCAM0	14.16	12.30	13.13

types of noise (babble, white, factory and volvo noise) extracted from the NOISEX-92 database (Varga and J.M.Steeneken, 1993). Noise was added to the clean signals at two SNR levels (10 dB and 15 dB) both for narrowband and wideband speech. The noise-corrupted test sets were decoded using the DNN and TDNN acoustic models trained using noisy speech. The obtained WER values are reported in Table 8 for the DNN acoustic model and in Table 9 for the TDNN acoustic model. It can be observed that the proposed method improved the recognition performance compared to the baseline at both SNR levels and for all four noise types both in narrowband and wideband speech. As already reported in Section 4.2 for clean speech, the recognition performance was again best when the proposed method was combined with the VTLN and SRA techniques. We conducted recognition experiments also with lower SNR values of 0 dB and 5 dB, but in these cases the proposed method was not able to show clear improvements compared to the baseline due to the deteriorated performance of LP to model formants when speech is severely distorted by noise.

5. Conclusion

Recognition of children's speech using ASR systems trained with adults' speech is subject to acoustic mismatch which leads to poor recognition performance. The main reason of the acoustic mismatch is the difference in the vocal tract length between adult and child speakers. This physiological difference results in largely distinct formant frequency values in speech signals produced by adult and child speakers. In order to tackle the effect of this acoustic mismatch in ASR, the current study proposes using an LP-based formant modification method to modify the formants of children's speech in the system test phase to be closer to those in adults' speech in the system training phase. The proposed method was used in ASR experiments by training the acoustic model using adults' speech from the WSJCAM0 database and by testing the system performance using children's speech from the PF_STAR database by processing the children's speech data with the proposed formant modification method. The results showed that the proposed method overcome the baseline (i.e., the ASR system where no processing of children's speech was adopted in testing) by a large margin. For the DNN and TDNN acoustic models, the proposed method gave relative WER improvements of 24% and 11%, respectively, for narrowband speech. For wideband speech, the corresponding relative improvements were 27% and 13%. We also compared the proposed method with two techniques (VTLN and SRA) which have previously been used in recognition of children's speech to reduce the effect of the acoustic mismatch between training and testing. Our experiments indicated that the proposed method gave clearly better performance compared to VTLN and SRA. The best performance in recognition of children's speech was achieved by combining the proposed method with VTLN and SRA yielding relative improvements of 30% and 22% compared to the baseline with the DNN and TDNN acoustic models, respectively, for narrowband speech. For wideband speech, the corresponding relative improvements were 14% and 29%. Furthermore,

Table 8
WERs obtained for noisy speech by the baseline system, by the proposed formant modification method and by combining the proposed method with VTLN and SRA. The experiments were conducted using the DNN acoustic model. Relative improvements of the proposed method and the combined system compared to the baseline are reported.

Noise type	Speech bandwidth	SNR (dB)	WER (%)			Relative improvement (%)	
			Baseline	Proposed	Proposed + VTLN + SRA	Proposed	Proposed + VTLN + SRA
Babble	Narrowband	10 dB	30.63	22.32	**17.02**	27.13	44.43
		15 dB	29.69	21.77	**16.19**	26.67	45.46
	Wideband	10 dB	30.22	19.41	**18.16**	35.77	39.90
		15 dB	26.34	17.52	**15.30**	33.48	41.91
White	Narrowband	10 dB	33.55	24.60	**20.12**	26.67	40.02
		15 dB	28.59	21.71	**16.50**	24.06	42.28
	Wideband	10 dB	24.20	16.15	**15.52**	33.26	35.86
		15 dB	21.33	14.78	**14.59**	30.70	31.59
Factory	Narrowband	10 dB	41.54	28.82	**20.69**	30.62	50.19
		15 dB	32.45	22.42	**16.89**	30.90	47.95
	Wideband	10 dB	30.82	18.87	**17.12**	38.77	44.46
		15 dB	26.20	16.90	**14.42**	35.58	44.45
Volvo	Narrowband	10 dB	22.53	17.15	**13.54**	23.87	39.90
		15 dB	22.00	18.21	**12.62**	17.22	42.63
	Wideband	10 dB	19.39	13.68	**12.75**	29.44	34.24
		15 dB	18.60	13.61	**12.42**	26.82	33.77

Table 9
WERs obtained for noisy speech by the baseline system, by the proposed formant modification method and by combining the proposed method with VTLN and SRA. The experiments were conducted using the TDNN acoustic model. Relative improvements of the proposed method and the combined system compared to the baseline are reported.

Noise type	Speech bandwidth	SNR (dB)	WER (%)			Relative improvement (%)	
			Baseline	Proposed	Proposed + VTLN + SRA	Proposed	Proposed + VTLN + SRA
Babble	Narrowband	10 dB	20.81	18.67	**16.31**	10.28	21.62
		15 dB	17.69	15.97	**15.17**	9.72	14.24
	Wideband	10 dB	16.60	14.89	**14.28**	10.30	13.97
		15 dB	14.43	13.38	**13.11**	7.27	9.14
White	Narrowband	10 dB	21.50	18.87	**17.92**	12.23	16.65
		15 dB	21.15	18.21	**15.69**	13.90	25.81
	Wideband	10 dB	17.33	16.65	**15.52**	3.92	10.44
		15 dB	15.30	14.88	**14.33**	2.74	6.33
Factory	Narrowband	10 dB	19.53	18.67	**17.88**	4.40	8.44
		15 dB	17.27	16.34	**15.76**	5.38	8.74
	Wideband	10 dB	16.66	14.54	**13.85**	12.72	16.86
		15 dB	14.56	13.97	**13.38**	4.05	8.10
Volvo	Narrowband	10 dB	19.57	16.42	**13.16**	16.09	32.24
		15 dB	17.75	15.85	**12.62**	10.70	28.90
	Wideband	10 dB	14.27	13.41	**11.79**	6.02	17.37
		15 dB	14.12	13.13	**11.55**	7.01	18.20

we tested the ASR system with noise-corrupted children's speech using additive noise of different types and SNR values and found that the proposed method gave improved WER values compared to the baseline both for the DNN and TDNN acoustic models when SNR was 10 dB and 15 dB. For more severely corrupted speech, however, the proposed method was not able to improve the performance.

In summary, the current study has shown that recognition of children's speech using ASR systems trained with adults' speech can be improved using the proposed LP-based formant modification method

and further improvements can be achieved by combining the formant modification method with two existing techniques (VTLN and SRA). In addition to yielding clearly improved WER values compared to the baseline, the proposed method is straightforward to implement. However, in order to improve recognition of children's speech which is severely corrupted by noise, the proposed method does not improve the ASR performance compared to the baseline and new research is needed. A potential topic of future research to address this issue is to use noise-robust LP methods, such as those studied in Airaksinen et al. (2018), Sambur and Jayant (1976), Pohjalainen et al. (2008) and Magi et al. (2009), instead of conventional LP to better model formants of noisy speech by the proposed method.

The proposed method showed consistent improvements compared to baselines in various scenarios (i.e. for wideband and narrowband speech, in noisy conditions, and using different acoustic models). The experiments were conducted, however, by exclusively using only one type of speech (read speech) from existing databases. We expect that the proposed method may not be affected by the type of speech and therefore we assume that the results may be transferable to spontaneous speech also. New experiments with spontaneous speech are, however, needed to get a better understanding about the performance of the proposed method for spontaneous speech. In addition, further studies are needed to understand how the performance of the proposed method compares to using an acoustic model trained using in-domain children's speech. Furthermore, in studying how the amount of training data of adult speech affects the system performance by testing the system using child speech of the PF_STAR database (as done in the current article), it would be appropriate to train the system using a large database of British English adult speech (such as the British subset from the Common Voice corpus (https://commonvoice.mozilla.org/en/datasets) instead of Librispeech. Moreover, for better acoustic models, future studies could be conducted by adding the PF_STAR child speech training set to the adult training set and run additional TDNN training iterations.

CRediT authorship contribution statement

Hemant Kumar Kathania: Conceptualization, Methodology, Software, Data curation, Validation, Data analysis, Interpretation, Investigation, Writing – review & editing. **Sudarsana Reddy Kadiri:** Conceptualization, Methodology, Software, Data curation, Validation, Data analysis, Interpretation, Investigation, Writing – review & editing. **Paavo Alku:** Methodology, Data analysis, Interpretation, Investigation, Review and editing, Resources, Supervision, Funding. **Mikko Kurimo:** Methodology, Data analysis, Interpretation, Investigation, Review and editing, Resources, Supervision, Funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Academy of Finland (grants 329267, 330139). The computational resources were provided by Aalto ScienceIT.

References

Ahmad, W., Shahnawazuddin, S., Kathania, H., Pradhan, G., Samaddar, A., 2017. Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion. In: Proc. INTERSPEECH 2017. pp. 2391–2395. http://dx.doi.org/10.21437/INTERSPEECH.2017-302.

Airaksinen, M., Juvela, L., Räsänen, O., Alku, P., 2018. Time-regularized linear prediction for noise-robust extraction of the spectral envelope of speech. In: INTERSPEECH. pp. 701–705.

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.

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.

Claes, T., Dologlou, I., ten Bosch, L., van Compernolle, D., 1998. A novel feature transformation for vocal tract length normalization in automatic speech recognition. IEEE Trans. Speech Audio Process. 6 (6), 549–557.

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.

Cosi, P., 2009. On the development of matched and mismatched Italian children's speech recognition system. In: Proc. INTERSPEECH, pp. 540–543.

Dahl, G., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Speech Audio Process. 20 (1), 30–42.

Dubagunta, S.P., Hande Kabil, S., Magimai.-Doss, M., 2019. Improving children speech recognition through feature learning from raw speech signal. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5736–5740.

Fainberg, J., Bell, P., Lincoln, M., Renals, S., 2016. Improving children's speech recognition through out-of-domain data augmentation. In: INTERSPEECH 2016. pp. 1598–1602. http://dx.doi.org/10.21437/INTERSPEECH.2016-1348.

Hinton, G., Deng, l., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 29, 82–97. http://dx.doi.org/10.1109/MSP.2012.2205597.

Huber, J., Stathopoulos, E., Curione, G., Ash, T., Johnson, K., 1999. Formants of children, women, and men: The effects of vocal intensity variation. J. Acoust. Soc. Am. 106, 1532–1542. http://dx.doi.org/10.1121/1.427150.

Kathania, H.K., Ahmad, W., Shahnawazuddin, S., Samaddar, A.B., 2018. Explicit pitch mapping for improved children's speech recognition. Circuits Systems Signal Process. 32, 2021–2044.

Kathania, H.K., Ghai, S., Sinha, R., 2013. Soft-weighting technique for robust children speech recognition under mismatched condition. In: 2013 Annual IEEE India Conference (INDICON). pp. 1–6.

Kathania, H.K., Kadiri, S.R., Alku, P., Kurimo, M., 2020. Study of formant modification for children ASR. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7429–7433.

Kathania, H.K., Shahnawazuddin, S., Adiga, N., Ahmad, W., 2018. Role of prosodic features on children's speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5519–5523.

Kathania, H.K., Shahnawazuddin, S., Ahmad, W., Adiga, N., 2019. Role of linear, mel and inverse-mel filterbanks in automatic recognition of speech from high-pitched speakers. Circuits Systems Signal Process. 38, 4667–4682.

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).

Kathania, H.K., Shahnawazuddin, S., Sinha, R., 2014. Exploring HLDA based transformation for reducing acoustic mismatch in context of children speech recognition. In: 2014 International Conference on Signal Processing and Communications (SPCOM). pp. 1–5.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., Belpaeme, T., 2017. Child speech recognition in human-robot interaction: Evaluations and recommendations. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. In: HRI '17, Association for Computing Machinery, New York, NY, USA, pp. 82–90. http://dx.doi.org/10.1145/2909824.3020229.

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. 3, IEEE, pp. III–349.

Lee, S., Potamianos, A., Narayanan, S.S., 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. 105 (3), 1455–1468.

Magi, C., Pohjalainen, J., Bäckström, T., Alku, P., 2009. Stabilised weighted linear prediction. Speech Commun. 51 (5), 401–411, [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639308001799.

Makhoul, J., 1975. Linear prediction: A tutorial review. Proc. IEEE 63 (4), 561–580.

Narayanan, S., Potamianos, A., 2002. Creating conversational interfaces for children. IEEE Trans. Speech Audio Process. 10 (2), 65–78.

Narayanan, S., Potamianos, A., 2002. Creating conversational interfaces for children. IEEE Trans. Speech Audio Process. 10 (2), 65–78.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210.

Peddinti, V., Povey, D., Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proc. INTERSPEECH, pp. 3214–3218.

Pohjalainen, J., Magi, C., Alku, P., 2008. Enhancing noise robustness in automatic speech recognition using stabilized weighted linear prediction (SWLP). In: INTERSPEECH.

Potamianos, A., Narayanan, S., 2003. Robust recognition of children's speech. IEEE Trans. Speech Audio Process. 11 (6), 603–616.

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: Proc. INTERSPEECH 2018. ISCA, pp. 3743–3747.

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: Proc. ASRU.

Rath, S.P., Povey, D., Veselý, K., Černocký, J., 2013. Improved feature processing for deep neural networks. In: Proc. INTERSPEECH.

Robinson, T., Fransen, 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.

Russell, M., D'Arcy, S., Qun, L., 2007. The effects of bandwidth reduction on human and computer recognition of children's speech. IEEE Signal Process. Lett. 14 (12), 1044–1046.

Sambur, M., Jayant, N., 1976. LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise. IEEE Trans. Acoust. Speech Signal Process. 24 (6), 488–494.

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. IEEE, pp. 55–59.

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M., Strope, B., 2010. Your word is my command: Google search by voice: A case study. In: Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics. pp. 61–90, ch. 4.

Scukanec, G.P., Petrosino, L., Squibb, K., 1991. Formant frequency characteristics of children, Young adult, and aged female speakers. Perceptual Motor Skills 73 (1), 203–208.

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.

Shahnawazuddin, S., Adiga, N., Kathania, H.K., 2017. Effect of prosody modification on children's ASR. IEEE Signal Process. Lett. 24 (11), 1749–1753.

Shahnawazuddin, S., Adiga, N., Kathania, H.K., Sai, B.T., 2020. Creating speaker independent ASR system through prosody modification based data augmentation. Pattern Recognit. Lett. 131, 213–218.

Shahnawazuddin, S., Dey, A., Sinha, R., 2016. Pitch-adaptive front-end features for robust children's ASR. In: INTERSPEECH.

Shahnawazuddin, S., Kathania, H.K., Dey, A., sinha, R., 2018. Improving children's mismatched ASR using structured low-rank feature projection. Speech Commun. 105, 103–113.

Shahnawazuddin, S., Kathania, H.K., Sinha, R., 2015. Enhancing the Recognition of Children's Speech on Acoustically Mismatched ASR System. In: Proc. TENCON.

Shivakumar, P.G., Georgiou, P., 2020. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. Comput. Speech Lang. (ISSN: 0885-2308) 63, 101077. http://dx.doi.org/10.1016/j.csl.2020.101077.

Smith, J.O., Abel, J.S., 1999. Bark and ERB bilinear transforms. IEEE Trans. Speech Audio Process. 7 (6), 697–708.

Strube, H.W., 1980. Linear prediction on a warped frequency scale. J. Acoust. Soc. Am. 68 (4), 1071–1076.

Varga, A., J.M.Steeneken, H., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.

Yadav, I.C., Shahnawazuddin, S., Govind, D., Pradhan, G., 2018. Spectral smoothing by variationalmode decomposition and its effect on noise and pitch robustness of ASR system. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5629–5633.

Yeung, G., Alwan, A., 2018. On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children. In: Proc. INTERSPEECH 2018.

Yildirim, S., Narayanan, S., Byrd, D., Khurana, S., 2003. Acoustic analysis of preschool children's speech. In: ICPhS-15. pp. 949–952.

Zhu, X., Beauregard, G.T., Wyse, L.L., 2007. Real-time signal estimation from modified short-time Fourier transform magnitude spectra. IEEE Trans. Audio, Speech Lang. Process. 15 (5), 1645–1653.