STUDY OF FORMANT MODIFICATION FOR CHILDREN ASR

Hemant Kumar Kathania, Sudarsana Reddy Kadiri, Paavo Alku and Mikko Kurimo

Department of Signal Processing and Acoustics, Aalto University, Finland

(heamnt.kathania, sudarsana.kadiri, paavo.alko, and mikko.kurimo)@aalto.fi

ABSTRACT

The performance of automatic speech recognition systems for children's speech is known to suffer from the large variation and mismatch in the acoustic and linguistic attributes between children's and adults' speech. One of the various identified sources of mismatch is the difference in formant frequencies between adults and children. In this paper, we propose a formant modification method to mitigate differences between adults' and children's speech and to improve the performance of ASR for children. The explored technique gives a relative 27% improvement in system performance compared to a hybrid DNN-HMM baseline. We also compare the system performance with related speaker adaptation methods like vocal tract length normalization (VTLN) and speaking rate adaptation (SRA) and find that the proposed method gives improvements over them, as well. Combining the proposed method with VTLN and SRA results in a further reduction of WER. We also found that the proposed method performs well even for noisy speech.

Index Terms— Children speech recognition, Formant modification, DNN

1. INTRODUCTION

Automatic speech recognition (ASR) for children speech is a challenging task specifically under mismatched and noisy conditions [1, 2]. Mismatched conditions correspond to training the system with adults' speech and testing it with children's speech. Most of the ASR systems available publicly work well with adults' speech but in the case of children's speech or in the case of low-SNR speech, the system performance collapses [3]. This is due to the vocal tract variability of children speech [4, 5] and noise corrupting the formants and other key features in speech. One more important issue for children ASR is the limited amount of publicly available speech data for child speakers [6, 7]. For adults speech, we have databases of more than 1000hrs of training data for ASR building, but for children speech, databases of only a few hours are available even in English. For all these reasons, it is necessary that ASR systems built for children are robust for various mismatched conditions.

In the past two decades, research in speech recognition has made tremendous progress. Consequently, a large number of speech-based user applications have been developed [3]. In such applications, the performance of the deployed ASR system is affected by several factors. One of them is the inter-speaker variability such as age, gender, accent, speaking-rate, and formant frequencies of the speakers present across training and test data sets. To impart robustness towards this variability, the ASR models are trained on a large amount of speech data collected from different speakers. In addition, techniques like feature-space maximum likelihood linear regression (fMLLR) [8] and vocal tract length normalization (VTLN) [9] are commonly included to adapt to the variation.

Many studies have explored the changes in formant frequencies with age [4, 10, 11, 12]. Formant frequencies F1, F2, and F3 have been found to be highest in children, decrease with increasing age [4, 11] and be lowest in adult men. 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 amount of change in formant frequencies are smaller between older age groups than between the younger ones.

In this paper, a linear predictive coding (LPC) -based formant modification technique is proposed to overcome the differences in ASR between adults and children speech. The method aims at reducing differences in formant frequencies between adults and children speech. MFCC features are computed after the formant modification. The study shows that the formant modification improved the system performance compared to the VTLN and SRA techniques in recognition of children speech under mismatched conditions. Further, we extended our study to children ASR in noisy conditions and even in this case found a reduction in WER.

2. FORMANT MODIFICATION

The difference in the vocal-tract dimension between adults and children is the major cause for mismatch in ASR. Figure 1 demonstrates differences in the formant structure between adult and child speakers for the vowels "IY" and "EI". The blue and green curve shows the LPC spectrum for an adult and child speaker, respectively. The red curve shows the LPC spectrum modified with the proposed method from the child speech.

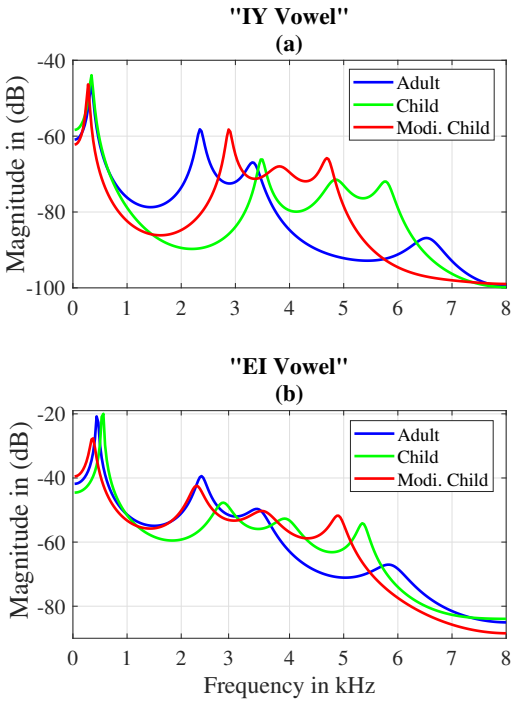

Fig. 1. LPC spectra computed from frames of the vowels /IY/ and /EI/ showing variations in formant frequencies. Blue and green curves were computed from speech utterances of an adult and child speaker, respectively. Red curve shows the spectrum after applying the formant modification for the spectrum of the child speaker.

Formant modification is carried out to the LP spectrum using warping. The resulting LP spectrum, denoted by $X_\alpha(f)$, is obtained by modifying the original LPC spectrum $X(f)$ by the warping function $w_\alpha(f)$, where α is the warping factor:

$$X_\alpha(f) = X(w_\alpha(f)). \tag{1}$$

In the classical LPC method, an estimate of the speech signal X(n) is obtained as a linear combination of the previous N sample values

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

and its Z-transform is given by

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

Here z^{-k} is the k unit delay filters, and a_k are the LPC filter coefficients using which the LPC spectrum can be computed. In this study, the unit delay filter is replaced by an all-pass filter D(z) to warp the LPC spectrum. The warping of the frequency scale is conducted using a first order all-pass filter [13, 14] given by

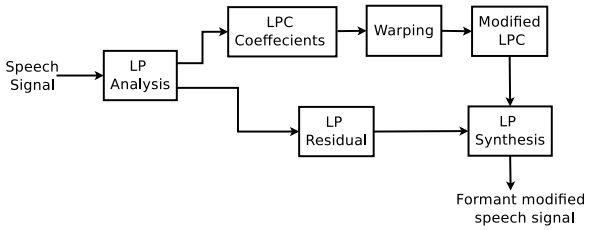

Fig. 2. A block diagram of the LPC based formant modification method

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

where α is a warping factor in the range of $-1 < \alpha < 1$. With a proper value of α, the warped frequency scale matches the psycho-acoustic scale based on auditory perception [15]. By applying the warping function D(z) to the LPC coefficients $a_k s$, the spectral resonances (formants) can be shifted systematically. With positive values of α, formant frequencies shift to lower frequencies as shown in Figure 1 by the red curves for "IY" and "EI" vowels. The modified LPC coefficients ($a_k^{'} s$) and the residual (x(n) - $\hat{x}(n)$) are used to synthesize the speech signal using a standard LPC synthesizer [16]. This synthesized signal is referred as the formant modified signal, and it is used in the current study as input to an ASR system. A schematic block diagram describing the steps involved in the proposed method is depicted in Figure 2.

3. SPEECH DATA AND EXPERIMENTAL SETUP

Two British English speech corpora, WSJCAM0 [17] and PF-STAR [18], were used in the experiments. The train set of WSJCAM0 has 92 adult (male and female) speakers and a total of 15.5 hours of speech data. For children speech, the train set of PF-STAR contains 8.3 hours of data from 122 speakers. The total duration of children speech data for testing is 1.1 hours. The age of the child speakers in this corpus varies between 4-14 years. The analyses were performed using wideband speech (sampled at 16 kHz).

For computing the MFCC feature vectors, speech data was first analyzed in overlapping Hamming-windowed 20-ms frames with a frame-shift of 10 ms. The 40-channel Mel-filterbank were used to compute 13-dimensional base MFCC features. The base MFCC features were then spliced in time, i.e., 4 frames to the left and to the right of the current analysis frame were appended making the feature vector dimension equal to 117. The dimensionality was reduced to 40 using linear discriminant analysis followed by maximum likelihood

linear transformation and de-correlation was performed with cepstral mean and variance normalization (CMVN). 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 [19].

The context-dependent hidden Markov model (HMM) was utilized to train the acoustic models. The observation probabilities for the HMM states were generated using the Gaussian mixture model (GMM) and DNN [20]. Cross-word triphone models consisting of a HMM with 8 diagonal covariance Gaussian components per state were used in the case of 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 once again considering a context size of 9. The number of hidden layers was chosen as 8 with each layer consisting of 1024 hidden nodes. The nonlinearity in the hidden layers was modeled using the *tanh* function. The initial leaning rate for training the DNN-HMM parameters was set at 0.015 which was reduced to 0.002 after 20 epochs and extra 10 epochs of training were employed. The minibatch size for neural network training was selected as 512.

To decode the children speech test set, a domain-specific bigram language model (LM) was used. This bigram LM was trained on the transcripts of the speech data of PF-STAR excluding the test set. The out-of-vocabulary (OOV) rate and perplexity of the bigram LM with respect to the children test set are 1.20% and 95.8, respectively. A lexicon of 1969 words including the pronunciation variations was employed.

4. RESULTS AND DISCUSSION

The baseline WERs in % for the children test set are given in Table 1. MFCC features are used for acoustic modeling. The acoustic and linguistic differences between the training (adults speech) and test (children speech) data degrade the recognition performances compared to matched cases [21, 22, 23, 24, 25]. Despite applying CMVN and fMLLR, the WERs are quite poor even in the case of DNN-HMM-based ASR systems. To overcome the formant frequency differences between adults and children speech, and to improve the recognition performance, the proposed formant modification method is applied. The formant modification algorithm has a tunable parameter α that was varied from 0.05 to 0.25 in order to modify formant frequencies. In Figure 3, WER is shown by varying the formant modification factor α with the GMM and DNN acoustic models. In Figure 3, the blue dotted line shows the baseline for the GMM acoustic model and the red dotted line shows the baseline for the DNN acoustic model. It is interesting to note that after applying the formant modification technique, the performance of the GMM model moves closer

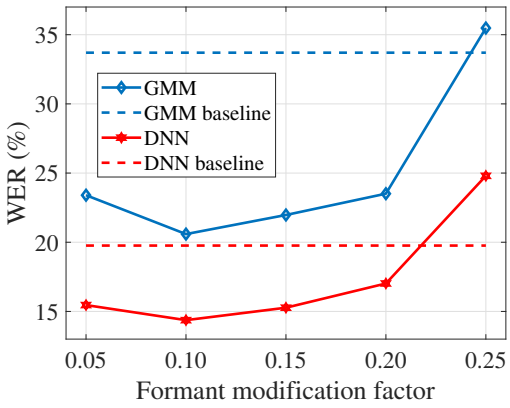

Fig. 3. WERs illustrating the effect of the formant modification factor (α) on the recognition of children speech using GMM and DNN-based ASR systems trained on adults speech.

Table 1. WERs for the children speech test set with respect to adult data trained ASR systems demonstrating the effect of the formant modification (Proposed) and two other methods (VTLN and SRA).

Acoustic model	WER (in %)				Rel. Imp. over baseline
	Baseline	VTLN	SRA	Proposed	
GMM	33.70	24.30	22.04	**20.59**	38.90
DNN	19.76	15.23	16.96	**14.37**	27.27

to the DNN baseline at $\alpha = 0.1$, and the DNN model further improved the performance with the formant modification.

WER values obtained by using the formant modification with the GMM and DNN acoustic models are reported in Table 1. It can be observed that the proposed method gives 33% and 19% relative improvements compared to the GMM and DNN based baseline ASR system, respectively. In Table 1, we have also compared our proposed method with VTLN and time scale modification based speaking rate adaptation (SRA) [26, 27] and found that the proposed method outperforms these two methods.

For further analysis, the children test data was divided into three different test sets based on three age groups ($4-6$ years, $7-9$ years, and $10-14$ years). In Table 2, baseline results for the age-wise test sets and average of all are shown. The proposed method improves the results in all the test sets.

To further enhance the system performance, we combined the proposed method with the SRA and VTLN techniques. The combinations studied are: proposed+VTLN, proposed+SRA, and proposed+ VTLN+SRA, and their results are reported in Table 3. Even though the combined

Table 2. WERs for the age-wise grouped children speech test sets with respect to adults data trained ASR systems demonstrating the effect of the formant modification (Proposed).

Age wise setup	WER (in %)		Relative Improvement
	Baseline	Proposed	
4 - 6	70.48	**49.69**	29.49
7 - 9	19.38	**10.69**	44.84
10 - 14	11.78	**10.53**	10.61
Avg	19.76	**14.37**	27.27

Table 3. WERs for the children speech test set with respect to adults data trained ASR systems demonstrating the effect of combining the proposed method with VTLN and SRA.

Acoustic model	WER (in %)				Rel. Imp. over pro. method
	Proposed	Proposed + VTLN	Proposed + SRA	Proposed + VTLN + SRA	
DNN	14.37	13.74	13.39	**12.35**	14.05

methods are better than the proposed methods alone, it seems that the proposed method provides some complementary information to VTLN and SRA. The best combination is proposed+SRA+VTLN which gives a relative improvement of 14% over the proposed method alone.

In order to further validate the effectiveness of the proposed formant modification method, another DNN-based ASR system was developed by pooling speech data from both the adults and children train sets. Such an ASR system reduces the degree of acoustic and linguistic mismatch by utilizing also children speech in the training. The baseline WER of the pooled system is given in Table 4. From Table 4, it can be seen that the proposed method and the combinations with the other techniques also reduce WER in the pooled system. A relative reduction of 11% in WER is noted compared to the baseline.

Table 4. WERs of the proposed method for the children speech test using an ASR system trained by pooling adults and children speech.

Acoustic model	WER in (%)			Rel. Imp. over baseline
	Baseline	Proposed	Pro. + VTLN + SRA	
DNN	12.26	11.25	**10.89**	11.17

To validate the robustness of the proposed method in noisy conditions, four different types of noise, viz. babble, white, factory and volvo noise extracted from NOISEX-92 [28], were added to the training and testing data under varying SNR values. The noisy test sets were decoded using the acoustic models trained with noisy speech. The WER for two different SNR values (10 dB and 15 dB) are reported in Table 5. It can be noted that for both SNR levels and for all the four different noise types, the performance of the proposed system is improved significantly. For lower SNR values, the proposed method did not improve the system performance. Further improvement can be seen for combinations with the other techniques (i.e, Proposed+VTLN+SRA).

Table 5. WERs of the proposed method for the children speech test set under varying additive noise conditions.

Noise Type	SNR (dB)	WER in (%)			Rel. Imp. over baseline
		Baseline	Proposed	Combination	
Babble	10dB	30.29	19.45	18.20	39.91
	15dB	26.46	17.59	15.34	42.02
White	10dB	24.11	16.09	15.52	35.62
	15dB	21.30	14.71	14.55	31.69
Factory	10dB	30.90	18.99	17.16	44.46
	15dB	26.30	16.94	14.47	44.98
Volvo	10dB	19.53	13.82	12.85	34.20
	15dB	18.72	13.70	12.69	32.21

5. CONCLUSION

In this paper, we have proposed and studied a formant modification method to demonstrate its effectiveness in the context of children speech recognition using acoustic models trained on adults speech. The proposed method gives a relative improvement of 27% over a baseline with DNN acoustic model using MFCC acoustic features. We have also compared the proposed method with the VTLN and SRA methods and found that the proposed method performs better. By combining the proposed method with SRA and VTLN, showed a further reduction in WER. A pooled system is also developed by pooling together speech data from both adult and children speakers and even in this case the proposed system manages to improve the performance. Further, we have developed and tested the ASR system using speech with additive noise, and for this case also found a reduction in WER.

6. REFERENCES

[1] A. Potaminaos and S. Narayanan, "Robust Recognition of Children Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, November 2003.

[2] P. Cosi, "On the development of matched and mismatched Italian children's speech recognition system," in *Proc. Interspeech*, 2009, pp. 540–543.

[3] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word

is my command: Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, 2010, ch. 4, pp. 61–90.

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

[5] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, February 2002.

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[7] F. Claus, H. Gamboa-Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about databases of children's speech," 2013, p. 2410–2414.

[8] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, 1995.

[9] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, January 1998.

[10] J. Huber, E. Stathopoulos, G. Curione, T. Ash, and K. Johnson, "Formants of children, women, and men: The effects of vocal intensity variation," *The Journal of the Acoustical Society of America*, vol. 106, pp. 1532–42, 10 1999.

[11] G. P. Scukanec, L. Petrosino, and K. Squibb, "Formant frequency characteristics of children, young adult, and aged female speakers," *Perceptual and Motor Skills*, vol. 73, no. 1, pp. 203–208, 1991.

[12] D. B. S. K. Serdar Yildirim, Shrikanth Narayanan, "Acoustic analysis of preschool children's speech," in *In ICPhS-15*, 2003, pp. 949–952.

[13] H. W. Strube, "Linear prediction on a warped frequency scale," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1071–1076, 1980.

[14] U. K. Laine, M. Karjalainen, and T. Altosaar, "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. IEEE, 1994, pp. III–349.

[15] J. O. Smith and J. S. Abel, "Bark and erb bilinear transforms," *IEEE Transactions on speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[16] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[17] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 1, May 1995, pp. 81–84.

[18] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. INTERSPEECH*, 2005, pp. 2761–2764.

[19] S. P. Rath, D. Povey, K. Veselý, and J. Černocký, "Improved feature processing for deep neural networks," in *Proc. INTERSPEECH*, 2013.

[20] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[21] S. Shahnawazuddin, H. Kathania, and R. Sinha, "Enhancing the recognition of children's speech on acoustically mismatched ASR system," in *Proc. TENCON*, 2015.

[22] H. K. Kathania, S. Shahnawazuddin, and R. Sinha, "Exploring hlda based transformation for reducing acoustic mismatch in context of children speech recognition," in *2014 International Conference on Signal Processing and Communications (SPCOM)*, July 2014, pp. 1–5.

[23] I. C. Yadav, S. Shahnawazuddin, D. Govind, and G. Pradhan, "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)*, April 2018, pp. 5629–5633.

[24] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's asr," in *INTERSPEECH*, 2016.

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

[26] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, July 2007.

[27] H. K. Kathania, S. Shahnawazuddin, W. Ahmad, N. Adiga, S. K. Jana, and A. B. Samaddar, "Improving children's speech recognition through time scale modification based speaking rate adaptation," in *2018 International Conference on Signal Processing and Communications (SPCOM)*, July 2018.

[28] "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 Communication*, vol. 12, no. 3, pp. 247–251, 21993.