Interspeech 2018
2-6 September 2018, Hyderabad

Non-Uniform Spectral Smoothing for Robust Children's Speech Recognition

Ishwar Chandra Yadav, Avinash Kumar, S. Shahnawazuddin and Gayadhar Pradhan

Department of Electronics and Communication Engineering
National Institute of Technology Patna, India.
{ishwarchy.ec15,k.avinash,s.syed,gdp}@nitp.ac.in

Abstract

Insufficient spectral smoothing during front-end speech parametrization results in pitch-induced distortions in the short-time magnitude spectra. This, in turn, degrades the performance of an automatic speech recognition (ASR) system for high-pitched speakers. Motivated by this fact, a non-uniform spectral smoothing algorithm is proposed in this paper in order to mitigate the acoustic mismatch resulting from pitch differences. In the proposed technique, the speech utterance is first segmented into vowel and non-vowel regions. The short-time magnitude spectrum obtained by discrete Fourier transform is then processed through a single-pole low-pass filter with different pole values for vowel and non-vowel regions. Sufficiently smoothed spectra is obtained by keeping higher values for the pole in the case of vowels while lower values are chosen for non-vowel regions. The Mel-frequency cepstral coefficients computed using the derived smoothed spectra are observed to be less affected by pitch variations. In order to validate this claim, an ASR system is developed on speech from adult speakers and evaluated on a test set which consists of children's speech to simulate large pitch differences. The experimental evaluations as well as signal domain analyses presented in this paper support the claim.
Index Terms: Children's speech recognition, spectral smoothing, pitch robust features, speech segmentation.

1. Introduction

The front-end parametric representation of acoustic data is an important phase in the development of any automatic speech recognition (ASR) system. It results in a compact representation of the input speech data by eliminating the irrelevant information and enhancing those aspects of the signal that contribute significantly towards the performance of the ASR systems. Earlier studies have reported that the most frequently used front-end parametrization techniques such as mel-frequency cepstral coefficients (MFCC) [1] are affected by the signal periodicity, especially for high-pitched children's speech [2, 3, 4]. Consequently, highly degraded performances have been reported when children's speech is transcribed using an ASR system trained on speech data from adult speakers. This is primarily due to insufficient smoothing of pitch harmonics during front-end feature extraction. In this paper, we present our attempts to deal with the issues arising from acoustic variability induced by age and gender differences, especially the pitch. A novel spectral smoothing technique is proposed to impart robustness towards pitch variations.

Acoustic attributes such as fundamental frequency (or pitch), formant frequencies and segmental durations vary with the age and gender of the speakers [5]. It is well known that, the intra- and inter-speaker spectral variability decreases as the age of the speaker increases [5, 6]. Generally, the fundamental frequency or pitch for adults' speech falls in the range of 80 Hz to 200 Hz while that for children's lies in the range of 200 Hz to 350 Hz. Elongation of vocal-tract occurs gradually as a child grows. This is accompanied with a decrement in formant frequencies [5, 7]. High pitch period in the case of children's speech creates widely spaced harmonic components due to under-sampling of the vocal-tract transfer function. Consequently, the probability of a harmonic component becomes more distal to the center frequency of a formant [8]. The increased spectral and temporal variability observed in the case of children's speech are primarily due to *anatomical and morphological differences in the vocal-tract geometry, less precise control of the articulators and a less refined ability to control suprasegmental aspects such as prosody* as highlighted in [9]. At the same time, compared to adults, their vocabulary is very limited and sometimes also contains some spurious words. As summarized in [10], children are more prone to use ungrammatical phrases, incorrect pronunciations and imaginative words.

Due to highly pronounced acoustic and linguistic variabilities in the case of children, automatic recognition of children's speech happens to be a much tougher problem [5, 9, 11]. To impart robustness towards acoustic variations, models are generally trained on a large amount of speech data collected from different groups of speakers. In addition to that, normalization techniques like feature-space maximum likelihood linear regression (fMLLR) [12] and vocal-tract length normalization (VTLN) [13] are included to mitigate the ill-effects resulting from age and gender variations. Despite that, higher pitch period leads to pitch-induced spectral distortions that affect the front-end speech parameterization process which, in turn, severely degrades the recognition performance. To address the pitch-induced spectral distortions, we present a novel spectral smoothing technique in this paper. The proposed approach employs *non-uniform single-pole filtering* to effectively smooth out the pitch harmonics from the spectra due to its low-pass characteristics. Therefore, by including the proposed approach for spectral smoothing into the standard MFCC feature computation process, pitch robustness is enhanced. The effectiveness of the proposed pitch robust features are experimentally evaluated in this paper. Furthermore, we have also studied the effect of combining the existing dominant feature-space normalization approaches with the proposed acoustic features.

The remainder of this paper is organized as follows: In Section 2, the proposed pitch-insensitive feature extraction approach is described. The experimental evaluations are presented in Section 3. Finally, the paper is concluded in Section 4.

2. Non-uniform spectral smoothing

2.1. Motivation for using non-uniform spectral smoothing

As already stated earlier, one of the most commonly used front-end acoustic features are MFCC. The MFCC features are designed to mimic the human perception mechanism. Due to low-

10.21437/Interspeech.2018-1828

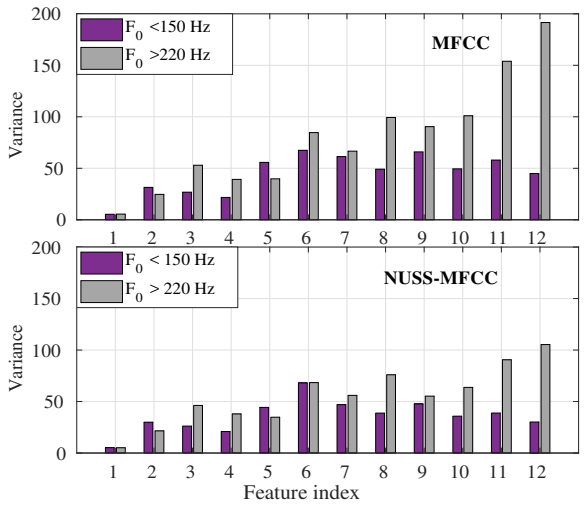

Figure 1: *Variances for the base MFCC features (C_1-C_{12}) for vowel /IY/ divided into two broad pitch (F_0) ranges. In the case of MFCC, the variance of higher-indexed cepstral coefficients for $F_0 > 220$ Hz range is high when compared to that for the $F_0 < 150$ Hz case (top panel). The variance mismatch for those coefficients is significantly reduced when the proposed spectral smoothing module based non-uniform low-pass filtering is included into the feature extraction process (bottom panel).*

time liftering of cepstral coefficients, the MFCC features are expected to be robust towards the ill-effects of excitation or signal periodicity. Contrary to that, the features vectors are observed to be affected by the signal periodicity especially when the speech signal being analyzed is from high-pitched (child) speakers [2, 4, 14]. Consequently, the MFCC features exhibit greater variance for the higher-indexed coefficients corresponding to speech data from high-pitched speakers in contrast to those for the speech data from low-pitched speakers [4]. To highlight this observation, we repeated the study on the variance of cepstral coefficients reported in [4] using voiced speech frames divided into two different pitch groups ($F_0 < 150$ HZ and $F_0 > 220$ Hz). Since reliable vowel markings are available in the TIMIT database, this analysis was performed on the vowel data extracted from the same. Using all the voiced speech frames belonging to a particular pitch group, the variance of each of the cepstral coefficients was computed. The results of that study are summarized in Figure 1. There is a significant mismatch in the variance of higher-indexed cepstral coefficients across the two pitch groups.

For high-pitched child speakers, ill-effects of pitch harmonics can be mitigated by reducing the length of the low-time lifter [14]. Even though this approach improves the recognition performance for child speakers, the performance with respect to adults' speech degrades. This is due to the loss of relevant spectral information when a large number of cepstral coefficients are removed by liftering. Motivated by that, in [3, 4], a spectral smoothing technique based on *pitch-adaptive cepstral truncation* (PACT) was proposed. In the case of PACT, first the short-time Fourier transform (STFT) employing a fixed duration Hamming window was performed to obtain the spectral representation of the speech signal. This was followed by deriving the log-compressed magnitude spectrum. Next, the cepstral coefficients were computed by applying inverse discrete Fourier transform (IDFT) on the log-compressed magnitude spectrum. A pitch-dependent low-time lifter was then used for smoothing out the pitch harmonics. The smoothed spectra was derived by transforming the liftered cepstrum back to the spectral domain using the discrete Fourier transform (DFT). This approach involves estimation of the mean pitch value for each of the utterances. The mean pitch value was, in turn, used for selecting the optimal lifter window length.

To be effective, PACT-based spectral smoothing relies on robust estimation of pitch of the signal being analyzed. Moreover, the same lifter is applied to all the frames irrespective of the fact whether the frame of speech being processed is voiced or unvoiced. Motivated by this issue, a non-uniform spectral smoothing is proposed in this work. In the proposed approach, we first segment the speech data into vowel and non-vowel like regions. The vowels are near-periodic, high energy and long duration sound units [15]. The non-vowels, on the other hand, are lower in magnitude. Further, the non-vowels like fricatives exhibit noise like characteristics and are shorter in duration. After speech segmentation, a single-pole low-pass filter is then used for spectral smoothing. Depending on the fact whether the speech frame being analyzed belongs to vowel or non-vowel like region, the pole location is changed. Thus, the proposed approach does not explicitly rely on pitch estimation. At the same time, the degree of applied spectral smoothing differs for vowel and non-vowel like regions.

2.2. Pitch smoothing through single-pole low-pass filter

In order to address the mismatch in the variances, we have explored non-uniform single-pole filtering for smoothing the spectra. The steps in the proposed scheme are as follows: The vowel like regions are first detected by using a recently reported method [16]. In that approach, an estimate of the speech signal at each time instant is obtained using non-local means (NLM) estimation [17]. Then the cumulative sum of the short-term magnitude spectrum is used as the front-end feature. The fluctuations in the feature is further smoothed by moving average filtering over a 50 ms window. The significant transitions in the smoothed feature are detected by convolving it with a 100 ms long first-order difference of Gaussian window having a standard deviation which is one-sixth of the window length. In the convolved output, termed as the *vowel detection evidence*, the peaks and valleys correspond to the vowel onset point (VOP) and vowel end point (VEP), respectively. The regions between them are selected as the vowels.

After segmenting speech data into vowel and non-vowel regions, STFT is performed with a fixed duration Hamming window. Thus the spectral representation of the speech signal is obtained. Based on the knowledge of vowel and non-vowel like regions, the magnitude spectra corresponding to each frame is then processed though a single-pole filter having different pole values for the vowel and non-vowel regions (α_V,α_{NV}). The transfer function for the single-pole filter is given as:

$$H(z) = \frac{1}{1 - \alpha z^{-1}}, \quad \text{where} \quad \alpha \in \{\alpha_V, \alpha_{NV}\} \quad (1)$$

MFCC features are then computed using the smoothed spectra. The block diagram for deriving the smoothed spectrum and the pitch-robust acoustic features is shown in Figure 2. The proposed acoustic features are referred to as NUSS-MFCC in the remaining of this paper.

The spectral smoothing obtained by the proposed approach is demonstrated using a set of spectral plots shown in Fig. 3. In Fig. 3 (a), the original log-compressed short-time magnitude

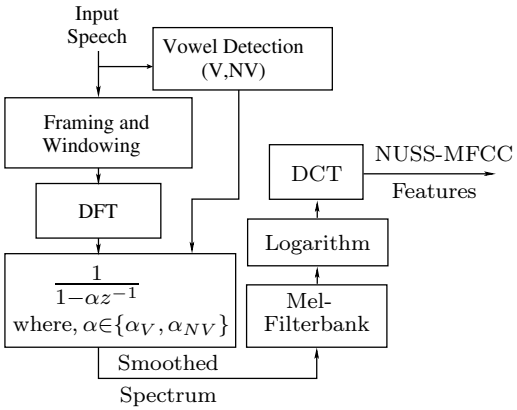

Figure 2: *Block diagram for computing the pitch-robust acoustic features applying non-uniform single-pole filtering for spectral smoothening.*

spectra for a voiced frame of speech having fundamental frequency $300Hz$ is shown. The spectra is then processed through a single-pole filter given by Eq. (1). The four pole values considered for deriving the smoothed spectra are 0.1, 0.4, 0.7 and 0.9 and the correspondingly smoothed spectra are shown in Fig 3 (b), (c), (d) and (e), respectively. It is evident from the shown spectral plots that, for high-pitch voiced speech units, lower values for pole do not result in sufficient spectral smoothing. At the same time, larger pole values lead to over smoothing. The desired amount of smoothing is obtained when the pole value is around 0.7. In this case, the smoothed spectra closely resembles the spectral envelope. As a consequence of sufficient spectral smoothing, the variance of the higher-indexed cepstral coefficients gets reduced which easily noticeable from Figure 1. In the case of speech, the ripples in the magnitude spectrum are mostly due to the excitation source information. The excitation source information is undesirable for ASR and, therefore, should be effectively removed. Spectral smoothing via the proposed method helps in removing the source information to a large extent, because of the low pass filtering effect.

3. Experimental evaluations

In this section, we present the results of the simulation studies done for evaluating the effectiveness of the proposed front-end acoustic features over the MFCC features.

3.1. Experimental setup

Overlapping Hamming windows of length 20 ms with frame-shift of 10 ms were employed to derive the short-time frames of speech. The 13-dimensional base MFCC features were extracted after warping the power spectra using a 40-channel Mel-filterbank. Time-splicing of the base MFCC features considering a context size of 9 was performed next. Dimensionality reduction and de-correlation were then done using linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT) to create 40 dimensional vectors. During spectral smoothing, the pole value α_V was varied from 0.5 to 0.8 in steps of 0.1. Similarly, α_{NV} was varied from 0.3 to 0.7 in steps of 0.1. The set of optimal values were chosen empirically. The window length, frame-rate and the number of channels in the Mel-filterbank were kept the same to derive the base NUSS-MFCC features. Time-splicing followed by LDA and MLLT were performed on the base NUSS-MFCC to obtain

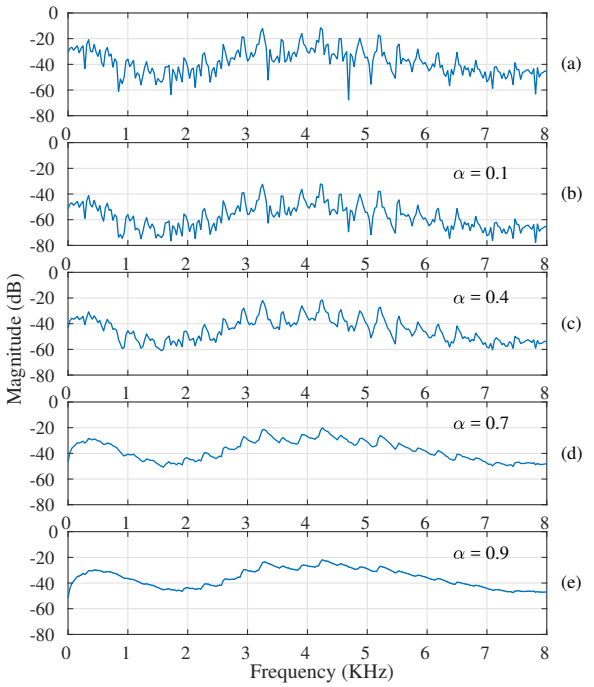

Figure 3: *Spectral smoothing affected by the proposed approach. (a) Log-compressed magnitude spectra for the central frame corresponding to voiced speech from a high-pitched speaker. The smoothed spectra obtained when (b)* $\alpha_V = 0.1$*, (c)* $\alpha_V = 0.4$*, (d)* $\alpha_V = 0.7$ *and (e)* $\alpha_V = 0.9$*.*

40 dimensional feature vectors. Both the kinds of feature vectors were subjected to cepstral mean and variance normalization (CMVN). Feature normalization using fMLLR was also done in order to reduce the ill-effects of speaker-dependent variations.

The ASR system used for evaluation was trained on the data obtained from the WSJCAM0 **British English** adults' speech corpus [18]. The Kaldi speech recognition toolkit [19] was used for the experimental studies presented in this paper. For statistically learning the ASR system parameters, a training set consisting of 15.5 hours of speech data from 92 adult male/female speakers was derived from WSJCAM0. The number of utterances in the training set was equal to $7,852$ with a total $132,778$ words. Context-dependent hidden Markov models (HMM) were employed for capturing the temporal variations. Initially, Gaussian mixture models (GMM) were used to generate the observation probabilities for the HMM states. Cross-word triphones were modeled using 3-states HMM each with 8 covariance components per state. Decision tree-based state tying, with the maximum number of senones being fixed at 2000, was performed. We have also explored acoustic modeling based on deep neural networks (DNN) [20] and Long Short-Term Memory (LSTM) [21] in this work. For DNN-HMM training, the fMLLR-normalized feature vectors were spliced in time once more using context size of 9 frames. The DNN-HMM setup consisted of 8 hidden layers. Each hidden layer was, in turn, composed of 1024 nodes employing *tanh* nonlinearity. The initial and final learning rates were selected to be 0.005 and 0.0005, respectively. A minibatch size of 512 was used while training the DNN parameters. The LSTM-based acoustic models were trained with 4 hidden layers each having 1024 nodes. The dimension of the LSTM cell was chosen as 1024. The number of epochs used for LSTM training was

Table 1: *WERs for children's speech development set with respect to DNN-based ASR system trained on adults' speech. The WERs demonstrate the effect of varying α_{NV} and α_V.*

α_{NV} \ α_V	WER (in %)				
	0.5	0.6	0.7	0.8	0.9
0.3	15.23	15.46	15.57	15.44	18.34
0.4	15.49	15.73	15.43	15.61	17.49
0.5	15.70	15.94	15.67	15.39	17.49
0.6	16.40	16.02	15.92	**14.65**	17.84
0.7	17.51	16.90	16.20	15.29	16.13
Baseline	16.62				

set to 5 while the initial and final learning rates were selected to be 0.005 and 0.0005, respectively. The initial alignments employed in DNN and LSTM training were obtained using the GMM-HMM system.

3.2. Evaluating pitch robustness of proposed features

Two different test sets were used for evaluation. The first one consisted of 0.6 hours of speech data from 20 adult speakers with a total of 5,608 words. The 5k MIT-Lincoln bi-gram language model (LM) was used while decoding this test set. This LM has perplexity of 95.3 with respect to the adults' speech test set. A lexicon comprised of 5,850 words including possible pronunciation variations was used. To simulate pitch-mismatched testing scenario, a development set and a test set consisting of children's speech were used. These sets were derived from the PF-STAR speech corpus (**British English**) [22]. The developmental set consisted of 150 utterances from 24 child speakers with a total 1.32 hours of speech data. The children's speech test set consisted of 1.1 hours of speech data from 60 child speakers with a total of 5,067 words. The child speakers in both development and test sets belonged to $3-14$ years age group. For decoding the children's speech test set, a 1.5k bigram LM trained on the transcripts of speech data in PF-STAR (excluding the test set) was used. This LM has an OOV rate of 1.20% and perplexity of 95.8 with respect to the children's speech test set. Moreover, a different lexicon consisting of 1,969 words was employed.

The WERs for the children's speech development set on adult data trained DNN-HMM-based ASR system with variation in the α_V and α_{NV} values are given in Table 1. The best combination is decided by the least possible WER as highlighted in the table. The WERs for both adults' and children's speech test sets are given in Table 2. The value of α_V and α_{NV} are chosen using the results given in Table 1 and the are 0.80 and 0.60, respectively. On comparing with the matched case testing (i.e., adults' test set), the recognition performances are extremely poor in the pitch-mismatched setup when MFCC features are used. Similar observations have been noted in earlier reported works as well [14, 2, 4]. The use of proposed features leads to significant reduction in WERs due to reduced pitch mismatch. When compared to PACT-MFCC, a relative improvement of $\mathbf{8.15}\%$ is obtained by using the proposed NUSS-MFCC features in the case of LSTM-based system.

Table 2: *WERs for the adults' and children's speech test sets with respect to ASR system trained on adults' speech data. Separate ASR systems are trained using MFCC, PACT-MFCC and NUSS-MFCC features.*

Test Corpus	Feature kind	WER (in %)		
		GMM	DNN	LSTM
Adult	MFCC	7.24	5.89	5.13
	PACT-MFCC	7.24	5.96	5.39
	NUSS-MFCC	7.26	5.97	5.20
Child	MFCC	33.52	19.27	16.33
	PACT-MFCC	31.27	17.20	15.94
	NUSS-MFCC	26.95	16.05	**14.64**

Table 3: *WERs for the children's speech test set with respect to adult data trained ASR systems demonstrating the effect of combining VTLN the proposed features. The percentage relative improvement (PRI) obtained by including VTLN are also given.*

Acoustic model	WER (in %)		P.R.I (%)
	fMLLR	VTLN+fMLLR	
GMM	26.95	19.86	26.30
DNN	16.05	13.92	12.90
LSTM	14.64	**12.79**	12.63

3.3. Reducing the ill-effects of formant scaling

Earlier works have shown that the use of VTLN is extremely effective in the case of children's ASR [3, 4, 23]. Hence, we have also explored VTLN to reduce the effect of formant scaling on the proposed NUSS-MFCC features. The linear frequency warping factors was varied from 0.70 to 1.12 in steps of 0.02. A maximum likelihood grid search under the constraints of the first-pass transcription was employed to select the optimal warping factor. The first-pass transcription was derived by decoding the unwarped features using the developed acoustic models. To obtain enhanced recognition performance, the optimally warped feature vectors were re-decoded. By the application of linear frequency warping, a large reduction in WER is obtained as evident from Table 3. It is worth mentioning here that, VTLN warped features were subjected to fMLLR transformation as well. This study shows that VTLN is additive with the proposed features.

4. Conclusion

A novel spectral smoothing technique to enhance the pitch robustness of front-end acoustic features is presented in this paper. The proposed approach involves two steps. First, the given speech data is segmented into vowel and non-vowel like regions. Next, the magnitude spectra corresponding to each of the short-time frames is processed using a single-pole filter. The pole location is changed depending on the fact whether the frame being analyzed belongs to vowel or non-vowel like region. The smoothed spectra thus obtained is used for computing the front-end acoustic features that are more robust towards pitch variations than the existing ones. This claim has been experimentally verified in this paper.

5. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[2] R. Sinha and S. Ghai, "On the use of pitch normalization for improving children's speech recognition." in *Proc. INTERSPEECH*, 2009, pp. 568–571.

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

[4] R. Sinha and S. Shahnawazuddin, "Assessment of pitch-adaptive front-end signal processing for childrens speech recognition," *Computer Speech & Language*, vol. 48, pp. 103 – 121, 2018.

[5] 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.

[6] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition," in *Proc. ICASSP*, vol. 1, May 2006, pp. I–I.

[7] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[8] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *JHSR*, vol. 9, pp. 421–447, 1976.

[9] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, November 2003.

[10] S. S. Gray, D. Willett, J. Pinto, J. Lu, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices," in *Proc. INTERSPEECH, Workshop on Child, Computer and Interaction*, 2014.

[11] 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.

[12] 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.

[13] 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.

[14] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Proc. INTERSPEECH*, 2009, pp. 1607–1610.

[15] K. N. Stevens, *Acoustic Phonetics*. The MIT Press Cambridge, Massachusetts, London, England, 2000.

[16] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Non-local estimation of speech signal for vowel onset point detection in varied environments," in *Proc. Interspeech 2017*, 2017, pp. 429–433.

[17] B. H. Tracey and E. L. Miller, "Nonlocal means denoising of ecg signals," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2383–2386, September 2012.

[18] 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.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. ASRU*, December 2011.

[20] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Magazine*, 2012.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] 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.

[23] R. Serizel and D. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, April 2016.