# SPECTRAL SMOOTHING BY VARIATIONAL MODE DECOMPOSITION AND ITS EFFECT ON NOISE AND PITCH ROBUSTNESS OF ASR SYSTEM

*Ishwar Chandra Yadav*[1], *S. Shahnawazuddin*[1], *D. Govind*[2] *and Gayadhar Pradhan*[1]

[1]Department of Electronics and Communication Engineering, NIT Patna, India
[2]Center for Computational Engineering and Networking, Amrita University, India
ishwarchy.ec15@nitp.ac.in, s.syed@nitp.ac.in, d_govind@cb.amrita.edu, gdp@nitp.ac.in

## ABSTRACT

A novel front-end speech parameterization technique that is robust towards ambient noise and pitch variations is proposed in this paper. In the proposed technique, the short-time magnitude spectrum obtained by discrete Fourier transform is first decomposed in several components using variational mode decomposition (VMD). For sufficiently smoothing the spectrum, the higher-order components are discarded. The smoothed spectrum is then obtained by reconstructing the spectrum using the first-two modes only. The Mel-frequency cepstral coefficients computed using the VMD-based smoothed spectra are observed to be affected less by ambient noise and pitch variations. To validate the same, an automatic speech recognition system is developed on clean speech from adult speakers and evaluated under noisy test conditions. Furthermore, experimental evaluations are also performed on another test set which consists of speech data from children to simulate large pitch differences. The experimental evaluations as well as signal domain analyses presented in this paper support these claims.

***Index Terms*****—** Speech recognition, ambient noise, pitch mismatch, spectral smoothing, VMD.

## 1. INTRODUCTION

Automatic speech recognition (ASR) is the process for converting speech signal captured as acoustic pressure waves into its corresponding sequence of words by means of computers. Earlier ASR applications included simple tasks such as voice dialing, interactive voice response, etc. Recent years have witnessed an exponential growth in computing power as well as amount of speech data available for system development. Consequently, ASR systems are being deployed in more challenging and complex user applications such as voice-based web search [1], interactions with hand-held mobile devices, etc. In such real-world applications, the ASR systems are exposed to varied operating conditions. People using smart phones inside cars, buses and trains is a very common sight. For effective operation, improving noise robustness of the employed ASR system is an important and challenging aspect. In addition to ambient noise, ASR systems employed in real-life applications are accessed by users of varying age and gender. Pitch (fundamental frequency) and formant frequencies are two such speaker-dependent acoustic attributes that vary with age and gender [2, 3]. Hence, such systems should also be robust towards the speaker-dependent variations (say pitch variations). To impart robustness towards speaker-dependent variations, acoustic models are generally trained on a large amount of speech data collected from different classes of speakers. Furthermore, techniques like feature-space maximum likelihood linear regression (fMLLR) [4] and/or vocal tract length normalization (VTLN) [5] are generally included to reduce the ill-effects of age and gender variations. Similarly, additional front-end speech processing modules are also included to mitigate the ill-effects of ambient noise [6].

In this paper, we present a novel front-end speech parameterization technique that simultaneously enhances the robustness towards ambient noise as well as pitch variations. The proposed approach is an extension of the dominant speech parameterization technique called Mel-frequency cepstral coefficients (MFCC) [7]. In our approach, a spectral smoothing module is added to the standard MFCC feature extraction process. In this regard, the short-time magnitude spectrum obtained by discrete Fourier transform (DFT) is decomposed into several components using variational mode decomposition (VMD) [8, 9]. Spectral smoothing is achieved by discarding the higher-order components and reconstructing the spectrum using the first-two modes. Acoustic features extracted using the smoothed spectra are observed to less sensitive to ambient noise and pitch variations. The experimental evaluations as well as signal domain analyses presented in this paper demonstrate same.

VMD has been employed in several tasks related to time-series analysis [10–13]. Some recent works have also explored VMD-based signal decomposition for speech analysis [14]. To the best of our knowledge, use of VMD-based spectral smoothing with application to ASR has not been reported yet. Further, unlike other reported techniques exploiting VMD algorithm, decomposition of spectrum is performed in this study. The rest of this paper is organized as follows: In Section 2, the proposed front-end speech parameterization technique employing spectral smoothing through VMD is described. In Section 3, the experimental evaluations demonstrating the effectiveness of the proposed features are presented. Finally, the paper is concluded in Section 4.

## 2. PROPOSED FRONT-END ACOUSTIC FEATURES

The steps involved in the proposed front-end feature extraction technique are summarized in Fig. 1. In addition to the usual steps in MFCC feature extraction process, a spectral smoothing module is included in the proposed approach. The intended spectral smoothing is achieved by decomposing the short-time magnitude spectra into several modes using VMD as mentioned earlier. The spectrum is then reconstructed using the first two modes only. MFCC features are then computed using the smoothed spectra. The proposed acoustic features are, therefore, referred to as VMD-MFCC in the remaining of this paper. The VMD-MFCC features are observed to be more robust towards noise as well as pitch variations. In the following subsection, a very brief introduction of VMD algorithm is

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

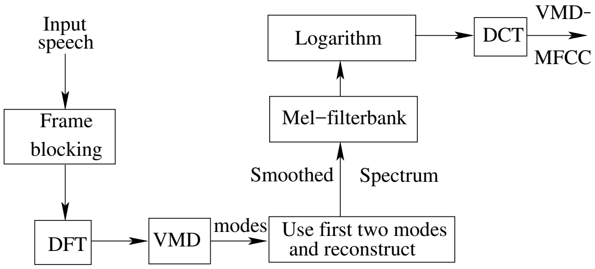

**Fig. 1:** Block diagram representing the steps involved in the extraction of the proposed front-end acoustic features employing VMD-based spectral smoothing.

presented. This is followed by a discussion on the effect of spectral smoothing in reducing the ill-effects of ambient noise and pitch periodicity.

### 2.1. Variational mode decomposition algorithm

Variational mode decomposition (VMD) is the technique to non-recursively decompose a sequence into a discrete number of bandlimited sub-sequences referred to as modes [8]. Each of the modes, in turn, has a compact frequency support around a center frequency. In order to identify these modes, a constrained optimization routine exploiting alternating direction method of multipliers is employed. During optimization step, the sum of the bandwidth of modes is minimized subject to the condition that the sum of the modes exactly reconstructs the original signal [8, 9]. In general, the number of modes is fixed before optimization.

### 2.2. Effect of spectral smoothing on noise and pitch

Since speech is a slowly varying non-stationary signal, spectral analysis of speech is done on short-time frames. Hence, magnitude spectra corresponding to each of the short-time frames is subjected to VMD-based spectral smoothing. To do so, the given magnitude spectrum is decomposed into several modes using the VMD algorithm. Next, the magnitude spectrum is reconstructed back using first few modes only. The spectral smoothing affected by this approach is demonstrated using the set of spectral plots shown in Fig. 2. In Fig. 2(a), the original log-compressed short-time magnitude spectrum for a voiced frame of speech signal is shown. This spectrum is decomposed into 8 modes and then reconstructed back after dropping the higher-order modes. The reconstructed magnitude spectrum by combining 4 lower-order modes is shown in Fig. 2(b). Similarly, the magnitude spectra derived by combining 3 and 2 lower-order modes are shown in Fig. 2(c) and Fig. 2(d), respectively. Finally, the magnitude spectrum obtained by retaining only the first mode is shown in Fig. 2(e).

It is evident from the shown spectral plots that, dropping higher-order modes leads to smoothing of the spectrum. Since further spectral smoothing will happen due to the use of Mel-filterbank and low-time lifter, dropping all higher-order modes and retaining only the first one may subsequently lead to over-smoothing. Therefore, this case was not considered for deriving the smoothed spectra. In this study, we have used the smoothed spectra obtained by combining the two lower-order modes only. The resulting spectrum, as clearly visible from Fig. 2(d), closely resembles the spectral envelope. The ripples in the magnitude spectrum, in the case of speech signal, are predominantly due to the excitation source information. In the context of ASR, excitation source information is undesirable and hence should be removed. Spectral smoothing via VMD helps in removing the source information to a large extent. Therefore, spectral smoothing is eventually expected to improve the recognition performance of the ASR system.

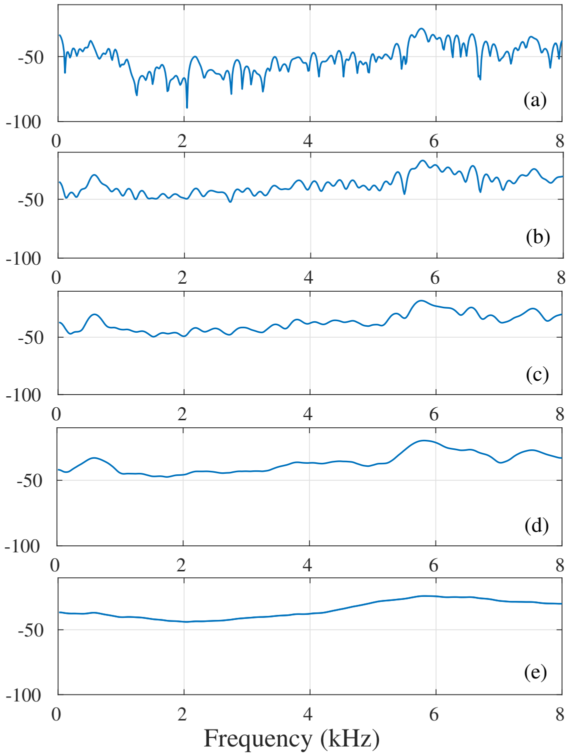

**Fig. 2:** (a) short-time magnitude spectrum for a frame of voiced speech. The reconstructed smoothed spectrum obtained by combining (b) first 4 modes, (c) first 3 modes, (d) first 2 modes and (e) first mode only. In each figure, x-axis represents the frequency in Hz and y-axis represents the magnitude in dB.

To demonstrate the effect of spectral smoothing on ambient noise and pitch variations, we performed the following study. We took two speech signals having the same word level transcription spoken by a high-pitched (female) and a low-pitched (male) speaker. The speech signals from the female and male speakers are shown in Fig. 3(a) and Fig. 3(b), respectively. The corresponding pitch contours are depicted in Fig. 3(c) and Fig. 3(d), respectively. The pitch contours were derived using the Wavesurfer toolkit [15]. Despite the context being same, pitch for speech data from female speaker is significantly higher than that for the male speaker. Next, both the speech samples were contaminated by adding 10dB noise. The speech waveforms corrupted by ambient noise are shown in Fig. 3(e) and Fig. 3(f) while their corresponding pitch contours are shown in Fig. 3(g) and Fig. 3(h), respectively. For each of the noise added speech files, the set short-time spectra was derived next. The magnitude spectra was then subjected to spectral smoothing using VMD as explained earlier. Finally, the speech signal was reconstructed from the smoothed spectra using overlap-add method after appending the

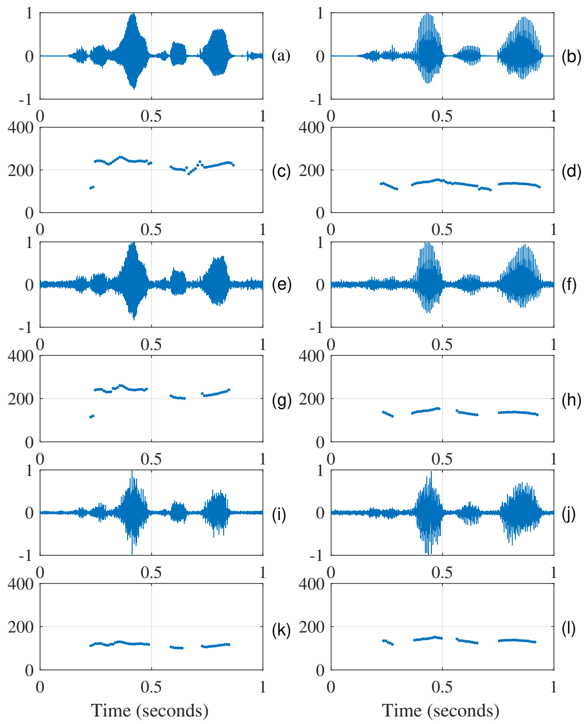

**Fig. 3:** (a) and (b) show a segment of clean speech data from female and male speakers, respectively. (c) and (d) the corresponding pitch contours for clean speech examples. (e) and (f) display the speech segments from female and male speakers corrupted by 10 dB noise, respectively, while the corresponding pitch contours are shown in (g) and (h), respectively. The reconstructed speech signals after applying VMD-based spectral smoothing are shown in (i) and (j), respectively, and their pitch contours are depicted in (k) and (l), respectively. In each figure, the x-axis represents time in seconds. For the speech signal, the y-axis denotes the amplitude. At the same time, y-axis represents the pitch in Hz for the pitch contours shown in this figure.

phase information. The reconstructed speech signals are shown in Fig. 3(i) and Fig. 3(j), respectively. Their pitch contours are shown in Fig. 3(k) and Fig. 3(l), respectively.

On comparing Fig. 3(e) and Fig. 3(f) with Fig. 3(i) and Fig. 3(j), the reduction in the noise can be easily noticed. At the same time, there is no significant change in the shape of the reconstructed signal when compared to the original clean waveforms. It may, therefore, be concluded that the proposed spectral smoothing reduces the ill-effects ambient noise. Similarly, on comparing the pitch contours for the original and reconstructed signals, significant reduction in pitch values for the female speaker is noted. On the other hand, the pitch contour for the male speaker remains almost the same. Thus, given that the linguistic context remains fixed, spectral smoothing results in similar pitch values for both male and female speakers. Reducing the pitch for female speakers and making it comparable to that for the male speakers is bound to reduce the pitch sensitivity of the ASR system. In the following section, we present the experimental evaluations that statistically validate the same.

## 3. EXPERIMENTAL EVALUATIONS

In this section, we present the results of the simulation studies done for evaluating the effectiveness of proposed front-end acoustic features over the MFCC features. First, the details of speech corpora and ASR system employed for evaluation are detailed. We then present the experimental evaluations under noisy test conditions. Finally, the simulation studies illustrating the effectiveness of proposed features under pitch-mismatched setup are presented.

### 3.1. Experimental setup

For computing MFCC features, overlapping Hamming windows of length 20 ms with frame-shift of 10 ms were employed to analyze speech data into short-time frames. In order to extract 13-dimensional base MFCC features, a 40-channel Mel-filterbank was used. The base MFCC features were then spliced in time considering a context size of 9 frames making the feature vector dimension equal to 117. Next, dimensionality reduction and de-correlation were performed using linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT) to obtain 40 dimensional feature vectors. The standard MATLAB code was used to perform VMD. The bandwidth constraint was chosen to be 2000 while the number of modes was fixed at 8. The values were selected through empirical studies. The window length, frame-rate and the number of channels in the Mel-filterbank were kept the same in the case of VMD-MFCC features as well. Furthermore, as in the case of MFCC, time-splicing followed by LDA and MLLT were performed on the base VMD-MFCC to obtain 40 dimensional feature vectors. Cepstral mean and variance normalization (CMVN) was applied to both the acoustic feature kinds. In addition to CMVN, feature normalization was also done using feature-space maximum likelihood linear regression (fMLLR) to boost the robustness towards speaker-dependent variations. The required fMLLR transformations were generated using speaker adaptive training [16].

The ASR system used for evaluation was trained on speech data obtained from the **British English** speech corpus WSJCAM0 [17]. The Kaldi toolkit [18] (accessed on June 2017) was used for all the experimental evaluations presented in this paper. For statistical modeling, a training set consisting of 15.5 hours of speech data from 92 adult male/female speakers was created from WSJCAM0. The number of utterances in the training set was equal to $7,852$ with a total $132,778$ words. For statistically learning the temporal variations, context-dependent hidden Markov models (HMM) were employed. Initially, the observation probabilities for the HMM states were generated using Gaussian mixture models (GMM). Cross-word triphone models consisting of 3-states HMM with 8 diagonal covariance Gaussian components per state were used for the GMM-HMM-based ASR system. Decision tree-based state tying was performed with the maximum number of senones being fixed at 2000.

After successfully implementing GMM-HMM system, acoustic modeling based on deep neural network (DNN) [19] was explored next. Before training DNN-HMM parameters, the fMLLR-normalized feature vectors were time-spliced once again considering a context size of 9 frames. The number of hidden layers in the DNN-HMM setup was fixed at 8 with each layer consisting of 1024 hidden nodes having *tanh* nonlinearity. An initial learning rate of 0.005 was chosen for training the DNN-HMM parameters. The learning rate was reduced to 0.0005 in 15 epochs. After reducing the learning rate, additional 5 epochs of training were employed. A minibatch

**Table 1:** WERs for the adults' speech test set with respect to GMM-HMM and DNN-HMM systems under clean and noisy test conditions demonstrating the effectiveness of proposed VMD-MFCC features over conventional MFCC.

| **Acoustic model** | **SNR dB** | **WER (in %)** | | **Relative imp. (%)** |
|---|---|---|---|---|
| | | MFCC | VMD-MFCC | |
| **GMM** | **Clean** | 7.24 | 7.31 | -0.9 |
| | **15** | 14.77 | 11.38 | 22.9 |
| | **10** | 31.09 | 23.01 | 25.9 |
| **DNN** | **Clean** | 5.89 | 6.01 | -2.0 |
| | **15** | 10.47 | 9.25 | 11.6 |
| | **10** | 23.09 | 19.62 | 15.0 |

size of $512$ was selected for neural net training. Furthermore, the initial state-level alignments employed in DNN training were generated using the earlier trained GMM-HMM system.

### 3.2. Evaluating noise robustness of the proposed features

To evaluate the noise robustness of the proposed acoustic features, a test set consisting $0.6$ hours of speech data was derived from the WSJCAM0 database. This test set comprised of data from 20 adult male/female speakers with a total of $5,608$ words. While decoding the adults' speech test set, MIT-Lincoln 5k Wall Street Journal bigram language model (LM) was used. This LM has a perplexity of $95.3$ with respect to the adults' test set while there are no out-of-vocabulary (OOV) words. A lexicon consisting of $5,850$ words including the pronunciation variations was used during decoding. The word error rate (WER) metric was used for evaluating the recognition performance.

The WERs for the adults' speech test set with respect to GMM- and DNN-based systems are enlisted in Table 1. The WERs are given for clean as well as noisy test conditions. For noisy testing, several different noises collected from NOISEX-92 database [20] such as factory noise, HF radio channel noise, pink noise, vehicle noise, engine room and operation room noises, etc., were added to the test data. The tabulated WERs are averaged over all the noise types. Furthermore, evaluations were performed for two different values of signal-to-noise ratio (SNR). Under clean testing scenario, both MFCC and proposed features result in almost similar WERs. On the other hand, VMD-MFCC features yield significantly lower WERs than MFCC when noise is added. The percentage relative improvements in WER are also tabulated to highlight the same.

### 3.3. Evaluating robustness towards pitch variations

Next, we evaluated the effectiveness of the proposed features under pitch-mismatched setup. To do so, another test set was derived from the PF-STAR speech corpus (**British English**) [21] consisting of speech data from child speakers. The employed 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 age of the child speakers in this test set lies in between $4-14$ years. While decoding the children's speech test sets, a domain-specific $1.5$k bigram LM was employed. This bigram LM was trained on the transcripts of speech data in PF-STAR excluding the test set. Further, a lexicon consisting of $1,969$ words including the pronunciation variations was employed. As stated earlier, children's speech is reported to have higher pitch (and also formant frequencies) when compared to adult male/female speakers due to anatomical differences. Furthermore, children's speech exhibits higher pitch variability than adults' speech. Consequently, transcribing children's speech on the developed acoustic models, leads to severe pitch mismatch.

**Table 2:** WERs for the children's speech test set with respect to GMM-HMM and DNN-HMM systems demonstrating the robustness of proposed VMD-MFCC features towards pitch variations.

| **Acoustic model** | **WER (in %)** | | **Relative imp. (%)** |
|---|---|---|---|
| | MFCC | VMD-MFCC | |
| **GMM** | 33.52 | 27.36 | 18.4 |
| **DNN** | 19.27 | 16.37 | 15.0 |

**Table 3:** WERs for the children's speech test set under noisy testing conditions demonstrating the effectiveness of proposed VMD-MFCC features over conventional MFCC.

| **Acoustic model** | **SNR dB** | **WER (in %)** | | **Relative imp. (%)** |
|---|---|---|---|---|
| | | MFCC | VMD-MFCC | |
| **GMM** | **15** | 57.55 | 49.28 | 14.2 |
| | **10** | 79.15 | 70.05 | 11.4 |
| **DNN** | **15** | 36.16 | 32.00 | 11.5 |
| | **10** | 65.07 | 56.65 | 12.1 |

The WERs for children's speech test set under clean conditions are given in Table 2. Compared to the matched case testing (adults' test set), severely degraded recognition performances are obtained in the mismatched setup. Similar observations have been noted in earlier reported works as well [22–24]. The anatomical differences among the two groups of speakers lead to acoustic mismatch which, in turn, results in higher WERs [25–27]. Several works have also been reported to improve the recognition of children's speech. The difference in pitch is one among the several factors contributing to the acoustic mismatch as highlighted in [28–32]. The use of proposed features leads to significant reduction in WERs due to reduced pitch mismatch. The relative improvements given in Table 2 obtained by using the proposed features highlight the same. Next, we evaluated the performance of the proposed features for children's speech recognition under noisy conditions. The WERs for that study are enlisted in Table 3. Like matched case testing, significant improvements are noted for noisy testing as well.

## 4. CONCLUSION

In the work presented in this paper, a novel front-end speech parameterization technique is presented. During the feature extraction process, the magnitude spectra is decomposed into several modes using variational mode decomposition. The smoothed spectra is obtained by dropping the higher-order modes prior to reconstruction. MFCC features are then computed using the smoothed spectra. The proposed acoustic features are observed to be more robust towards ambient noise as well pitch variations when compared to the conventional MFCC features. The signal domain analyses as well as the experimental evaluations presented in this study validate the same.

## 5. REFERENCES

[1] Johan Schalkwyk, Doug Beeferman, Franoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, chapter 4, pp. 61–90. 2010.

[2] S. Eguchi and I. J. Hirsh, "Development of speech sounds in children.," *Acta oto-laryngologica. Supplementum*, vol. 257, pp. 1–51, 1969.

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

[4] 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.

[5] Li 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.

[6] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, April 2014.

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

[8] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, February 2014.

[9] Soman K.P., Prabaharan Poornachandran, Athira S., and Harikumar K., "Recursive variational mode decomposition algorithm for real time power signal decomposition," *Procedia Technology*, vol. 21, no. Supplement C, pp. 540 – 546, 2015.

[10] Y. J. Xue, J. X. Cao, D. X. Wang, H. K. Du, and Y. Yao, "Application of the variational-mode decomposition for seismic time-frequency analysis," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3821–3831, August 2016.

[11] Ahmet Mert, "ECG feature extraction based on the bandwidth properties of variational mode decomposition," *Physiological Measurement*, vol. 37, no. 4, pp. 530, 2016.

[12] G. Jyothish Lal, E. A. Gopalakrishnan, and D. Govind, "Accurate estimation of glottal closure instants and glottal opening instants from electroglottographic signal using variational mode decomposition," *Circuits, Systems, and Signal Processing*, May 2017.

[13] P. D. Achlerkar, S. R. Samantaray, and M. S. Manikandan, "Variational mode decomposition and decision tree based detection and classification of powerquality disturbances in grid-connected distributed generation system," *IEEE Transactions on Smart Grid*, 2017.

[14] Abhay Upadhyay and Ram Bilas Pachori, "Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition," *Journal of the Franklin Institute*, vol. 352, no. 7, pp. 2679 – 2707, 2015.

[15] Kåre Sjölander and Jonas Beskow, "Wavesurfer - an open source speech tool," in *INTERSPEECH*, 2000, pp. 464–467.

[16] Shakti P. Rath, Daniel Povey, Karel Veselý, and January Černocký, "Improved feature processing for deep neural networks," in *Proc. INTERSPEECH*, 2013.

[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*, May 1995, vol. 1, pp. 81–84.

[18] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, December 2011.

[19] Geoffrey E. Hinton, Li Deng, Dong Yu, George Dahl, Abdel Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[20] "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, 1993.

[21] 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.

[22] Hank Liao, Golan Pundak, Olivier Siohan, Melissa K. Carroll, Noah Coccaro, Qi-Ming Jiang, Tara N. Sainath, Andrew W. Senior, Françoise Beaufays, and Michiel Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proc. INTERSPEECH*, 2015, pp. 1611–1615.

[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.

[24] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition," in *Proc. Spoken Language Technology Workshop (SLT)*, December 2014, pp. 135–140.

[25] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proc. Speech and Language Technologies in Education (SLaTE)*, September 2007.

[26] Sungbok Lee, Alexandros Potamianos, and Shrikanth S. Narayanan, "Acoustics of childrens 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.

[27] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction*, 2009, pp. 7:1–7:8.

[28] 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.

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

[30] Shweta Ghai and Rohit Sinha, "A study on the effect of pitch on LPCC and PLPC features for children's ASR in comparison to MFCC," in *Proc. INTERSPEECH*, 2011, pp. 2589–2592.

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

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