ROLE OF PROSODIC FEATURES ON CHILDREN'S SPEECH RECOGNITION

Hemant K. Kathania[1]*, S. Shahnawazuddin*[2]*, Nagaraj Adiga*[3] *and Waquar Ahmad*[1]

[1]Department of Electronics and Communication Engineering, NIT Sikkim, India
[2]Department of Electronics and Communication Engineering, NIT Patna, India
[3]Department of Computer Science, University of Crete, Greece
hemant.ece@nitsikkim.ac.in, s.syed@nitp.ac.in, nagaraj@csd.uoc.gr, waquar@nitsikkim.ac.in

ABSTRACT

In this paper, we have explored the role of combining prosodic variables with the existing acoustic features in the context of children's speech recognition using acoustic models trained on adults' speech. The explored acoustic features are Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction cepstral coefficients (PLPCC) while the considered prosodic variables are loudness, voice-intensity and voice-probability. An analysis presented in this paper shows that, given that the textual content remains the same, the considered prosodic variables exhibit very similar contours for adults' and children's speech. At the same time, the contours differ a lot when the context is different. Consequently, inclusion of prosodic information reduces the inter-speaker differences and increases the class discrimination. This subsequently improves the recognition performance. Further improvements are obtained by projecting the feature vectors obtained by combining the two features to a lower-dimensional subspace. The same has been experimentally verified in this study for mismatched speech recognition using deep neural network (DNN) based system. On combining MFCC (PLPCC) and prosodic features, a relative improvement of 16% (14%) is noted on decoding children's speech using adult data trained DNN models.

Index Terms— Children's ASR, acoustic mismatch, prosodic variables, feature projection.

1. INTRODUCTION

The primary objective of speech production and perception mechanism is to convey messages through a sequence of legal sound units. The intelligibility of spoken message is enhanced by including the information from melody, timing and stress in speech. This aspect enables the listener to segment continuous speech into phrases and words with ease [1]. Furthermore, speech signal also conveys many more lexical and non-lexical information such as tone, prominence, accent and emotion. The characteristics of the speech signal that enable humans to perceive these effects are collectively referred to as prosody. Prosody is concerned more with those elements of speech that reflect the properties of syllables and larger units of speech. A human listener can better recognize more familiar speakers in comparison to relatively less familiar ones due to speaker-specific prosody and the idiosyncrasies that are recognized by the listener [2]. Prosody has been studied as a knowledge source for speech understanding and has been explored in several tasks related to speech processing. Inclusion of prosodic information has been effectively used for language identification [3–5], text-to-speech (TTS) and voice conversion systems [6, 7] and automatic speech recognition (ASR) [8–10].

In this paper, we have explored the role of prosodic feature in the context of *children's mismatched ASR*. The task of recognizing children's speech using acoustic models trained on speech data from adult speakers is referred to as the mismatched ASR in this work. Earlier works have reported highly degraded recognition performances in the case of mismatched recognition tasks. The observed degradations are mainly due to large differences in both the acoustic and the linguistic correlates between the speech from adult and child speakers [11–14]. Several studies have been reported for addressing the acoustic mismatch in the context of children's mismatched ASR. Recently, a number of works have also explored acoustic modeling based on deep neural network (DNN) for improving children's speech recognition [15–19]. Yet, to the best of our knowledge, the role of prosodic features has not been explored in the context of children's mismatched ASR employing DNN-based acoustic modeling. The experimental evaluations presented in this study explore the effectiveness of combining acoustic features like Mel-frequency cepstral coefficient (MFCC) [20] and perceptual linear prediction coefficients (PLPCC) [21] with the prosodic variables in the context of children's mismatched ASR. For contrast, we have also explored the effectiveness of combining prosodic features with MFCCs/PLPCCs in those cases where the acoustic models are trained on speech data collected from both adult and children speaker. In both the cases, significant improvements are observed by the inclusion of prosodic information.

The rest of this paper is organized as follows: In Section 2, the motivation for using prosodic features is presented. The experimental studies demonstrating the effectiveness of including prosodic features in the context of children's speech recognition are presented in Section 3. Finally, the paper is concluded in Section 4.

2. MOTIVATION FOR USING PROSODY FEATURES

The prosodic variables explored in this work are loudness (LD), voice-probability (VP) and voice-intensity (VI). The prosody features used in this study are extracted using openSMILE [22] toolkit following procedure outlined in [23]. As a preliminary study, inclusion of prosodic information is first explored on the connected digit recognition task. This is followed by an analysis to justify why the inclusion of prosodic information helps in digit recognition.

2.1. Connected digit recognition

The training and the test data for the connected digit recognition was obtained from the TIDIGITS database [24]. This speech corpus contains 11.3 hours of speech data from 326 speakers (225 adults and 101 children). Each of the speakers utter one to seven digits

Table 1: The WERs for the GMM-HMM-based connected digit recognition system trained using MFCC features. The WERs obtained with respect to GMM-HMM system trained after frame-level concatenation of MFCC and prosody features are also given.

Test set	WER (in %)	
	MFCC	MFCC+Prosody
Adult	1.65	1.91
Child	9.17	6.93

long strings consisting of eleven different digits (0-9 and 'OH'). The age of the adult speakers contributing to this database varies from 17 to 70 years. The child speakers, on the other hand, belong to an age group of 6 to 15 years. A train set comprising of 5.3 hours of speech data from 197 adult male/female speaker was created from this database. Two different test sets were derived for testing. The first test set was composed of 1.6 hours speech data from 81 adult speakers. The other test set comprised of 1.9 hours speech data from 49 children. The speech data used for the connected digit recognition task was sampled at 8 kHz rate.

In order to evaluate the effectiveness of including prosodic information, an ASR system was developed on the adult data using the Kaldi speech recognition toolkit [25]. For extracting the front-end acoustic features, speech data was first analyzed using overlapping Hamming windows of length 20 ms with frame rate of 100 Hz. A 23 channel Mel-filterbank was employed to compute the 13-dimensional base MFCC features. This was followed by time-splicing of the base features considering a context size of 9, i.e., ± 4 frames. The dimensionality of the resulting time-spliced features was then reduced to 40 using linear discriminant analysis (LDA) [26]. Further de-correlation of the feature vectors was done through maximum likelihood linear transform (MLLT) [27]. Mean and variance normalization (MVN) was also performed. In order to develop the required classifier, the 11 digits (0-9 and 'OH') were modeled as whole words using continuous density hidden Markov models (HMM) employing 3 states per word including silence. Each HMM state, in turn, was modeled using 6 diagonal-covariance Gaussian mixture model (GMM). An equilikely wordnet was employed during testing. The metric used to measure the recognition performance is word error rate (WER).

The WERs for the connected digit recognition system are given in Table 1. It is to note that, only adults' speech training data was used for learning the GMM-HMM parameters. Therefore, the recognition performance for adults' speech test set is much better than that for children's speech. As highlighted earlier, the acoustic attributes of speech from adult and children speakers differ significantly. Hence, the observed differences in recognition performance are very much obvious. Similar degradation in performance was noted in earlier works as well [28–30]. Similar degradation is observed when systems trained on children's speech are employed for decoding the adults' speech data.

To study the effect of including prosodic information, another digit recognition system was retrained on adults' data after appending the prosodic and acoustic features. For extracting the prosodic variables, the speech data was analyzed into overlapping frames using Hamming window of length 20 ms with frame rate of 100 Hz. The 3-dimensional prosodic features were computed using the openSMILE [22] toolkit. Next, the prosodic variables and base MFCC feature vectors were concatenated at the frame level making the base feature dimension equal to 16. This was followed by time-splicing, dimensionality reduction and de-correlation. The final feature vector dimension after LDA+MLLT was chosen as 40. Mean and variance normalization was also applied to the feature vectors. The specifications of the GMM-HMM architecture was the same as explained earlier. The WERs for the adults' and children's speech test sets after concatenating MFCC and prosodic features are given in Table 1. The WER for children's speech is observed to reduce significantly when prosodic information is included.

Table 2: WERs for the connected digit recognition task with respect to children's speech test set demonstrating the effect of including prosodic information along with acoustic features.

WER (in %)					
Digit	MFCC	MFCC + Prosody	Digit	MFCC	MFCC + Prosody
One	5.40	7.10	Six	6.00	**3.80**
Two	4.90	**1.60**	Seven	9.60	**5.60**
Three	7.90	**7.20**	Eight	18.80	20.60
Four	6.30	**4.40**	Nine	1.00	**0.50**
Five	16.00	**5.40**	Oh	5.40	**4.00**

For children's speech test set, the WER for each of the eleven digits with and without the inclusion of prosodic variables are enlisted separately in Table 2. The corresponding WERs for the case when only MFCC features are used are also tabulated for proper contrast. As evident from the table, the WER is noted to reduce for most of the digits by including prosodic variables. For majority of the digits, the reduction in WER is very large. The WERs for all those cases have been presented in bold. On the other hand, the WER is found to increase for digits *one* and *eight*. To summarize, appending prosodic variables with MFCC is somehow increasing the discrimination among the classes thereby reducing the WER. In the following sub-section, we present an analysis which attempts to justify these observations.

2.2. Analyzing the cause of improved recognition performance

In order to develop ASR systems, given the training data, the relevant front-end features are extracted first. The front-end features along with the class-labels are then used to learn the statistical parameters. The performance depends on quality of the front-end features as well as the employed statistical modeling approach. If the front-end features are such that the discrimination among the classes is more, then better performance will be obtained. In other words, the front-end features should be such that the *within-class differences are minimal*. On the other hand, the *between-class differences are large*. Using this reasoning, we studied the nature of explored prosodic variables for each of the classes. The results tabulated earlier suggest that, for those digits where the WER had decreased, the nature should be similar for a given class irrespective of the speaker. In other words, for any given digit (class), the deviation in the value of the prosodic variables for adult and child speakers should be very small. At the same time, the nature should be dissimilar for those classes where adding prosodic information did not help.

In order to study the nature of any particular prosodic variable,

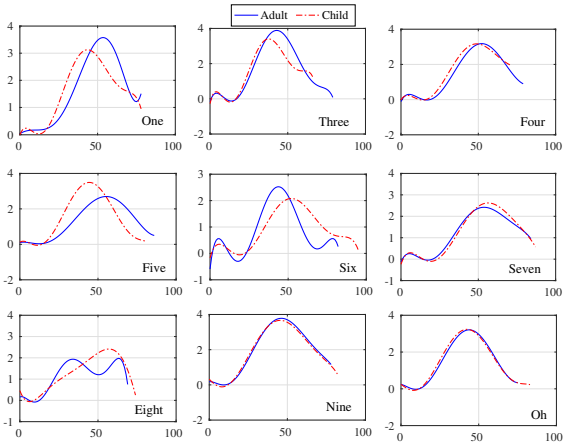

Fig. 1: Smoothed contours depicting the mean loudness for few of the digits. The x-axis depicts the frame index while the y-axis represents the mean value.

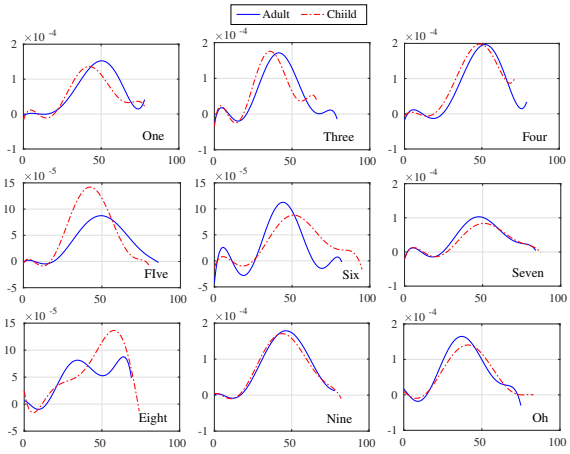

Fig. 2: Smoothed contours depicting the mean intensity for few of the digits. The x-axis depicts the frame index while the y-axis represents the mean value.

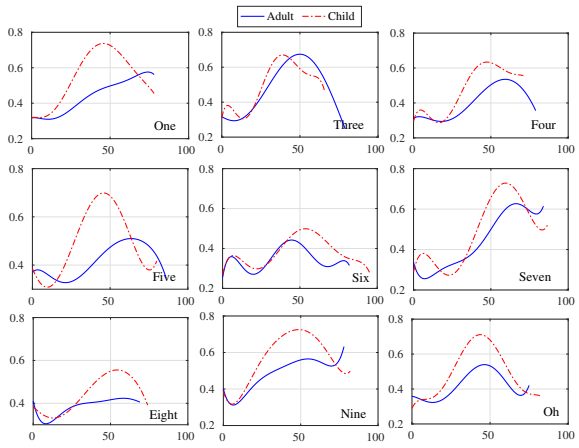

Fig. 3: Smoothed contours depicting the mean voice-probability for few of the digits. The x-axis depicts the frame index while the y-axis represents the mean value.

smooth contours were derived for each of the classes as follows. Forty isolated utterances of a given class, say digit one, were selected at random from the database. Twenty of those were collected from adults while remaining twenty were from child speakers. Next, the prosodic variables were computed for each of the utterances from the adult speakers. The mean value for each of the frames was then computed using all the twenty examples. In a similar manner, the mean was computed using the examples from the children. Finally, a seventh-order polynomial function was fitted over the mean data to derive a smooth contour. These steps were repeated for each of the classes and each of the prosodic variable kinds.

The smooth contours for loudness are shown in Fig. 1. For each of the classes, the blue (solid) curves are for the case when data from adult speakers is used. The red (dash-dot) curves are obtained using data from children. The plots are shown for few of the digits only due to lack of space. It is to note that, for digits *three*, *four*, *five*, *six*, *seven*, *nine* and *oh*, the contours look very similar for both adult and child speakers in each case. On the other hand, the contours for digits *one* and *eight* for adult speakers are starkly different from those for the children. The smooth contours for the remaining two prosodic variables are shown in Fig. 2 and Fig. 3, respectively. Observations similar to those stated in the case of loudness are evident in these cases as well. Referring to Fig. 1 - 3 and Table 2, the following three observations are worth highlighting:

- The smooth contours for the mean of the explored prosodic variables derived using adult and children speech happen to be very similar for those cases where the WERs have reduced.
- The WER had increased significantly for digits *one* and *eight* and the contours for adult and child speakers are also starkly different in those cases.
- For two different digits, the prosodic contours do not look similar thereby enhancing the inter-class differences.

Thus it may be concluded that using prosodic features reduces the within-class differences while enhancing the between-class differences. Therefore the recognition performance for children's speech with respect to adult data trained digit recognizer improves significantly. Motivated by these results, the role of prosodic variables in the context of continuous speech recognition especially under mismatched setup was explored next.

3. CONTINUOUS SPEECH RECOGNITION

In this section, we study the role of prosodic features in the context of continuous speech recognition task.

3.1. Experimental setup

For experimental evaluations, ASR systems were developed on the 15.5 hours adults' speech data from WSJCAM0 British English speech corpus [31] using the Kaldi toolkit. There are a total of 7861 utterances from 92 adult (male/female) speakers with approximately 90 sentences per speaker in this train set. For mismatched testing, the children's speech test set of the PF-STAR British English speech database [32] was employed. This test set contains 1.1 hours of

Table 3: WERs for the children's speech test set with respect to acoustic models trained on adults' speech. The WERs are given for the cases when MFCC as well as PLPCC features are used to train the GMM-HMM- and DNN-HMM-based ASR systems. The WERs are also tabulated for the cases when the prosodic variables are combined with the MFCC and PLPCC features.

Explored Acoustic Model	Acoustic Feature Kind	WER (in %)		
		Baseline	+ Prosody	+ Prosody + HLDA
GMM	MFCC	32.69	25.81	**20.63**
	PLPCC	33.21	26.96	**21.35**
DNN	MFCC	19.68	16.66	**12.73**
	PLPCC	20.16	17.51	**13.28**

speech data from 60 child speakers with a total of 5067 words. The experimental evaluations were performed on wideband speech.

For computing the MFCC/PLPCC feature vectors, the earlier described steps were followed with a difference that a 40-channel Mel-filterbank was used. Furthermore, to boost the robustness towards speaker variations, feature-space maximum likelihood linear regression (fMLLR) was employed for normalization. The observation probabilities for the HMM states were generated using the GMM and deep neural network (DNN) [33]. Cross-word triphone models consisting of a 3-states HMM with 8 diagonal covariance Gaussian components per state were used in the case of GMM-HMM-based ASR system. Further, decision tree-based state tying was performed with the maximum number of tied-states (senones) being fixed at 2000. While 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 learning 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 net training was selected as 512.

For decoding the children's speech test set, a domain-specific 1.5k bigram language model (LM) was employed. This bigram LM was trained on the transcripts of the speech data in PF-STAR excluding test set i.e., on the transcripts for the train set only. The out-of-vocabulary (OOV) rate and perplexity of the employed bigram LM with respect to the children's test set are 1.20% and 95.8, respectively. A lexicon of 1,969 words including the pronunciation variations was employed.

3.2. Evaluation results

The baseline WERs for children's test set with respect to the GMM-HMM- and DNN-HMM-based ASR systems trained using MFCC and PLPCC features, respectively, are given in Table 3. On combining the considered prosodic variables with either of the explored acoustic features, significant reductions in WERs are noted similar to that observed in the case of digit recognition task.

Its a common practice to apply some kind of dimensionality reduction and de-correlation technique whenever two different kinds of feature vectors are concatenated. This helps in reducing the redundancies and retaining relevant information. Moreover, projecting the data to a lower dimensional subspace is reported to be highly effective for children's speech recognition under mismatched setup [34, 35]. Motivated by this, heteroscedastic linear discriminant analysis (HLDA) was employed for learning a low-rank feature projection matrix. The 16-dimensional base features obtained by concatenating the MFCC/PLCC features with prosodic variables were used for deriving the HLDA matrix. The projection matrix was learned on the training data and then applied to both the train as well as the test sets. The base features obtained after low-rank projection were then spliced in time considering a context size of 9. This was followed by LDA and MLLT and processing via MVN and fMLLR. The so obtained feature vectors were then used for learning the parameters of the ASR system. On projecting the data to a lower-dimensional subspace, a significant reduction in WER was obtained. The best case WERs for those studies are given in Table 3. A relative reduction of around $10 - 12\%$ over the prosody included baseline was obtained by low-rank feature projection.

Table 4: WERs for children's speech test set with respect to the DNN-HMM-based ASR systems trained after pooling speech data from adult as well as child speakers.

Acoustic Model	WER (in %)		
	Baseline	+ Prosody	+ Prosody + HLDA
DNN	11.47	9.98	8.82

In order to further validate the effectiveness of combining the prosodic features with the acoustic features, another DNN-HMM-based ASR system was developed by pooling together speech data from both adult as well as children train sets. The children's speech train set derived from PF-STAR consisted of 8.3 hours of speech data from 122 children. The total number of utterances in this train set was equal to 856 with a total of 46,974 words. The developed ASR system exhibits a lower degree of acoustic/linguistic mismatch due to the pooling of children's speech into training. As a result, the baseline WERs for the developed system (given in Table 4) are observed to be significantly lower when compared to those obtained with respect to the ASR system trained on adults' speech only (see Table 3). Despite these facts, further reductions in WERs are noted when the prosodic features are combined with the acoustic features as given in Table 4. Further reductions in WER are obtained by projecting the features to lower-dimensional subspace using HLDA.

4. CONCLUSION

In this paper we have studied the effectiveness of combining prosodic features with two of the dominant acoustic features in the context of children's speech recognition using acoustic models trained on adults' speech. In such cases, significant degradation in recognition performance is noted due to a severe mismatch in the acoustic/linguistic attributes of the speech data from adult and child speakers. The prosodic variables explored in this paper are voice-probability, voice-intensity and loudness. On combining the prosodic variables with MFCC/PLPCC features, significant reductions in WERs are noted. To further explore the effectiveness of prosodic features on children's speech recognition, another ASR system is developed using speech data from both adult and child speakers. Even in this case, significant improvements are reported. In order to further improve the system performance, low-rank feature projection is also explored. Additive reductions in WERs are obtained by low-rank feature projection.

5. REFERENCES

[1] E. Shriberg, A. Stolcke, Hakkani-Tur, and G. D., Tur, "Prosody-based automatic segmentation of speech into sentences and topicss," *Speech Commununication*, vol. 32, pp. 127–154, 2001.

[2] G. Doddington, "Speaker recognition based on idiolectic differences between speakers," in *Proc. EUROSPEECH, Aalborg*, 2001, pp. 2521–2524.

[3] M. Komatsu, K. Mori, T. Arai, and Y. Murahara, "Human language identification with reduced segmental information: comparison between monolinguals and bilinguals," in *Proc. EUROSPEECH*, 2001, pp. 149–152.

[4] K. Mori, N. Toba, T. Harada, T. Arai, M. Komatsu, M. Aoyagi, and Y. Murahara, "Human language identification with reduced spectral information." in *Proc. EUROSPEECH*, 1999, pp. 391–394.

[5] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, no. 10, pp. 782–796, 2008.

[6] M. Ross, "Is syntactic structure prosodically retrievable?" in *Proc. European Conference on Speech Communication and Technology*, 1997.

[7] J. P. H. V. Santen, "Prosodic modeling in text-to-speech synthesis," in *Proc. European Conference on Speech Communication and Technology*, 1997.

[8] D. H. Milone and A. J. Rubio, "Prosodic and accentual information for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 321–333, July 2003.

[9] K. Hirose and K. Iwano, "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition," in *Proc. IEEE ICASSP, Istanbul*, 2000, pp. 1763–1766.

[10] K. Hirose, N. Minematsu, , and M. Ito, "Experimental study on the role of prosodic features in the human process of spoken word perception," in *Proc. ESCA Workshop on Prosody*, 2003.

[11] 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.

[12] S. Lee, A. Potamianos, and S. 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.

[13] 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.

[14] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction*, 2009, pp. 7:1–7:8.

[15] 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.

[16] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q. Jiang, T. N. Sainath, A. W. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proc. INTERSPEECH*, 2015, pp. 1611–1615.

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

[18] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's ASR," in *Proc. INTERSPEECH*, 2016, pp. 3459–3463.

[19] S. Shahnawazuddin, K. T. Deepak, G. Pradhan, and R. Sinha, "Enhancing noise and pitch robustness of children's ASR," in *Proc. ICASSP*, March 2017, pp. 5225–5229.

[20] 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.

[21] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 57, no. 4, pp. 1738–52, April 1990.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[23] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, 1st ed. Springer Publishing Company, Incorporated, 2016.

[24] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, 1984, pp. 42.11.1–42.11.4.

[25] 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.

[26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[27] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models." *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[28] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, vol. 1, May 1996, pp. 349–352.

[29] D. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. ICSLP*, vol. 2, October 1996, pp. 1145–1148.

[30] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proc. ICASSP*, vol. 1, May 1998, pp. 433–436.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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