ENHANCING NOISE AND PITCH ROBUSTNESS OF CHILDREN'S ASR

S Shahnawazuddin[1], Deepak K. T.[2], Gayadhar Pradhan[1] and Rohit Sinha[3]

[1]Department of Electronics and Communication Engineering, NIT Patna, India
[2]Department of Electronics and Communication Engineering, IIIT Dharwad, India
[3]Department of Electronics and Electrical Engineering, IIT Guwahati, India
s.syed@nitp.ac.in, deepak.thotappa@gmail.com, gdp@nitp.ac.in, rsinha@iitg.ernet.in

ABSTRACT

It is well known that, when noisy speech is transcribed using automatic speech recognition (ASR) systems trained on clean data, a highly degraded recognition performance is obtained. The problem gets further aggravated when the targeted group happens to be child speakers. For children's speech, the acoustic correlates such as pitch and formant frequency vary significantly with age. This makes the recognition of children's speech very challenging. In this paper, we have explored the ways to enhance the noise robustness of ASR systems for children's speech. Towards addressing the same, recently developed front-end acoustic features based on spectral moments (SMAC) are explored. The SMAC features are reported to be more noise robust than the conventional features like the mel-frequency cepsatral coefficients. At the same time, the SMAC features are also noted to be sensitive to the variations in the pitch. To reduce the pitch sensitivity, a spectral smoothing approach based on adaptive-liftering is proposed. Spectral smoothening prior to the computation of spectral moments results in a significant improvement in the robustness to pitch without affecting the noise immunity. To further enhance noise robustness, a foreground speech segmentation and enhancement module is also included in the proposed front-end speech parameterization technique.

Index Terms— Children's speech recognition, spectral smoothening, speech enhancement.

1. INTRODUCTION

The development of an automatic speech recognition (ASR) system comprises of two major parts viz. the acoustic and the linguistic aspects. The linguistic aspects deal with the creation of lexicon and the training of domain-specific language model (LM). The acoustic part can further be broken down to front-end speech parameterization and training of statistical acoustic models. For the last few decades, the hidden Markov model (HMM) has been the most widely used technique for learning the acoustic model parameters. With the recent developments in ASR, the observation probabilities for the states of the HMM are now being generated through the deep neural network (DNN) [1]. The front-end speech parameterization, on the other hand, deals with the task of deriving a compact representation of the raw speech. These short-time parametric representations are also referred to as the acoustic feature vectors. The chosen parametric representation intends to capture the relevant information in speech signal while removing the redundancies. The Mel-frequency cepstral coefficients (MFCC) and the perceptual linear prediction cepstral coefficients (PLPCC) are the two dominant examples of the commonly used ones. Speech recognition systems developed using the MFCC/ PLPCC features have been explored for a large number of speech-based applications.

The recognition performance of ASR systems are affected by a number of factors. One of the factor is the differences in the level of ambient noise in the training and test data. Severe performance degradation is noted when an ASR system trained on clean speech is tested using noisy speech. To address this shortcoming a noise robust front-end feature extraction approach based on spectral moments was proposed in [2]. Another factor affecting those features is the variations in the pitch of the speech signal used for training and testing. An extreme example of pitch mismatched ASR is the task of recognizing children's speech using acoustic models trained on adults' data and vice-versa. Both the acoustic and the linguistic correlates for children's speech differ from that for adults' [3, 4, 5]. In earlier works [6, 7], several front-end features such as the MFCCs, the linear prediction cepstral coefficient (LPCC), the PLPCC and the perceptual minimum variance distortionless response (PMVDR) were analyzed and were found to be sensitive to the variation in the average pitch values. A number of works have been reported to address the pitch-induced distortions [8, 9].

In this paper, we propose a front-end speech parameterization technique that is robust to both noise and pitch variations. To simulate the same, an ASR system is trained on speech data collected from both adult and child speakers while testing is done using clean as well as noisy speech from children. In order to enhance the noise robustness of the developed ASR system, acoustic features based on the first central spectral moment time-frequency distribution (also known as

SMAC) are explored. In the work reported in [2], the SMAC features were noted to be less affected by additive noise compared to the MFCCs and PLPCCs. At the same time, these features were noted to be sensitive to the pitch periodicity of the signal being analyzed. Consequently, we introduce a spectral smoothening step prior to the computation of spectral moments for enhancing the pitch robustness. This is done via pitch-adaptive-liftering of the cepstral coefficients [10]. In order to further boost the noise robustness, a foreground speech segmentation and enhancement module is included before the computation of feature vectors [11]. The combination of the two techniques is noted to significantly enhance the recognition performance of children's speech.

The rest of this paper is organized as follows: In Section 2, the proposed front-end speech parameterization approach is discussed. In Section 3, we present the evaluation of the proposed scheme. Finally, the paper is concluded in Section 4.

2. PROPOSED SPEECH PARAMETERIZATION APPROACH

A front-end speech parameterization approach based on the normalized first central Spectral Moment time-frequency distribution Augmented by low-order Cepstral coefficients (SMAC) was proposed recently in [2]. In that approach, the spectral moment components were computed from the speech spectra filtered using a set of mel-spaced Gabor filters. Invoking the notion of *pyknogram* [12], it was argued that the information about the resonances of the speech signal could be captured by the spectral moments. The pyknogram is a density plot of the frequencies present in the input signal. It is to note that the pyknogram does not model the relative importance of each resonance peak. Consequently, the low-order cepstral coefficients capturing the spectral slope were also appended to the spectral moments to derive the final feature vector. The experimental evaluations done under different noise conditions showed that the use of the SMAC features was superior to those of MFCC/RASTA-PLP features.

In practical conditions, ASR systems get exposed to not only the ambient noise but also to speakers of varying age and gender. The studies reported in [2] were performed on the *matched* ASR task, i.e., transcribing adults' speech using acoustic models trained on speech from adult speakers only. In such cases, the pitch-induced distortions are not that severe. On the other hand, in the case of *children's mismatched* ASR systems, effective smoothing of pitch harmonics in the spectra is required. In this paper, mismatched ASR refers the task of decoding children's speech on acoustic models trained either using the adults' speech only or by pooling speech data from both adult as well as child speakers. As already mentioned, a large difference exists between the pitch values for adult and child speakers. Even among the children themselves, the pitch variation with age of the speaker is more diverse than that for the case of adult speakers. These differences lead to the pitch-induced distortions that severely affect the performance of the mismatched ASR task [10]. Our experimental exploration revealed that, like the MFCCs and PLPCCs, even the SMAC features are sensitive to pitch-induced distortions.

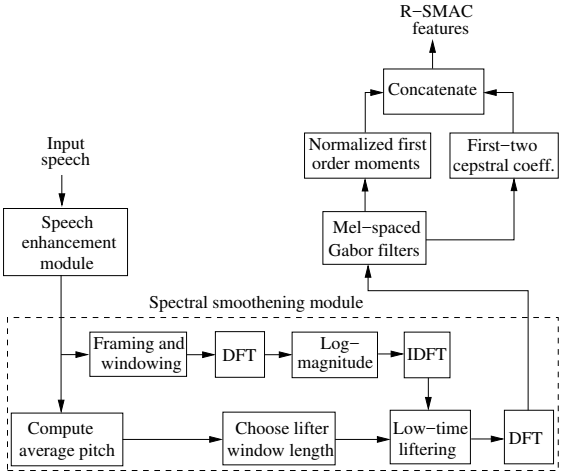

Fig. 1. Block diagram of proposed front-end speech parameterization approach employing foreground speech segmentation and enhancement as well as adaptive-liftering-based spectral smoothening to enhance noise and pitch robustness.

To address the aforementioned issues, a novel front-end speech parameterization approach is proposed in this paper. The block diagram of the proposed front-end speech parameterization technique is shown in Fig. 1. The proposed approach consists of two extra modules added to the SMAC feature extraction process in order to enhance the noise and pitch robustness. The resulting feature are referred to as *robust* SMAC (R-SMAC) features in this work. In the following, we discuss the adaptive-liftering-based spectral smoothening technique explored to address the pitch-induced distortions. This followed by a discussion on the foreground speech segmentation and enhancement scheme to further boost noise robustness.

2.1. Pitch-adaptive-liftering for spectral smoothening

In order to improve the pitch robustness, the proposed speech parameterization technique includes a spectral smoothening module based on adaptive-liftering of cesptral coefficients before computing the spectral moments. The steps in the proposed scheme are as follows: Using short-time Fourier transform (STFT) analysis involving a fixed duration Hamming window, the spectral representation of the speech signal is obtained. Next, the log-compressed magnitude spectrum is derived for each short-time frame of speech. This is followed by conversion to the cepstral representation using inverse discrete Fourier transform (IDFT). It is to note that the cepstral domain representation retains the periodicity of the speech excitation since the discussed steps are essentially equivalent to

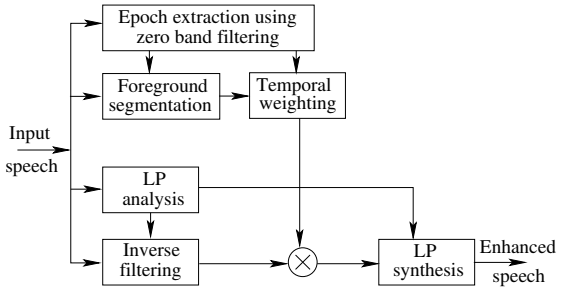

Fig. 2. Block diagram of the foreground speech segmentation and enhancement module.

the linear filtering. Consequently, a suitable low-time lifter is applied to smooth out the pitch harmonics. For determining the duration of the applied low-time lifter ℓ, the average pitch value f_0 for the utterance being analyzed is computed, such that $\ell = f_s/f_0$ where f_s is the sampling frequency. The average pitch value can be computed using any of the several approaches reported in literature viz. TEMPO [13], RAPT [14] and WaveSurfer [15]. The liftered cepstrum is then transformed back to the spectral domain using forward DFT. Given the smoothed spectrum, the front-end features based on spectral moments are derived.

2.2. Foreground speech segmentation and enhancement

In most of the speech-based applications, speech signal recorded in the natural environment gets contaminated by other interfering sources. The degradation of recorded speech signal impacts the quality and hence there is a necessity to enhance it. However, the interfering sources are not stationary in nature and their characteristics vary with respect to time. The interfering background noise can temporally overlap with the desired speech or it can be temporally isolated event in the recorded signal. It is therefore necessary to identify the desired foreground speech regions from rest of the background noise. Subsequently, the foreground speech regions can be enhanced further to improve the speech quality. In our recent work [11], a two stage approach was proposed which segments the foreground speech from rest of the background noise and subsequently enhances it. The block diagram of the enhancement module is shown in Fig. 2. In this work, the foreground speech segmentation and enhancement is used as a front-end pre-processing module to enhance the quality of the recorded speech signal.

3. EXPERIMENTAL EVALUATION

3.1. Experimental setup

For all the experimental evaluations reported in this study, two different British English speech corpora namely, WSJCAM0 [16] and PF-STAR [17] are used. The WSJCAM0 database consists of 15.5 hours of speech from 92 adult male/female speakers for training. In the training set of WSJCAM0 database, there are a total of 7861 utterances with approximately 90 sentences per speaker. On the other hand, the train set of PF-STAR contains 8.3 hours of data from 122 child speakers. The adults' speech test set available with WSJCAM0, consisting of 0.6 hours of speech data from 20 speakers, is used for the matched case testing. For mismatched testing, the test set in PF-STAR consisting of 1.1 hours of speech data from 60 child speakers is used. The children's speech test set consists of a total of 5067 words. All experimental evaluations are performed for the narrowband (sampled at 8 kHz rate) speech.

For the extraction of MFCC features, a Hamming window of length 20 ms with frame rate of 100 Hz and a pre-emphasis factor of 0.97 is used for speech data analysis. Using 23-channel Mel-filterbank, the 13-dimensional base MFCC features are computed. Next, the base MFCC features are time-spliced considering a context size of 9. This is followed by dimensionality reduction and decorrelation of the obtained feature vectors using linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT). The reduced dimensionality of the feature vector is chosen to be 40. For the extraction of SMAC features, a frame size of 20 ms with an overlap of 10 ms and a pre-emphasis factor of 0.97 is considered. A 12-channel mel-spaced Gabor filterbank with each filter having bandwidth of 237 mel is employed as suggested in [2]. The derived spectral moments are appended with first two cepstral coefficients (C_0, C_1) of the Gabor filtered spectra. This is followed by the post-processing of the base features using LDA and MLLT to yield the 40-dimensional SMAC features. Finally, the cepstral mean and variance normalization is applied to all studied acoustic features. In order to further improve the performance, both the kinds of features are normalized using the feature-space maximum likelihood linear regression (fMLLR). The required fMLLR transformations are generated for the training and test data using speaker adaptive training (SAT) approach. To optimize the fMLLR-transform, the SAT is performed on another ASR system employing Gaussian mixture models (GMM) for generating observation probabilities for the HMM states.

In the experimental evaluations presented in [2], the SMAC features were evaluated in the context of GMM-HMM-based acoustic modeling. Whereas, in this work, we use more advanced acoustic modeling framework based on DNN. It is to note that only a few works on the children's ASR employing DNN-based acoustic modeling have been reported [18, 19, 20]. The train set of both the aforementioned databases is pooled for learning the statistical parameters of the DNN-HMM-based ASR system. The ASR system development and testing is performed using the Kaldi toolkit [21]. The hidden layers in the DNN-HMM system included the *tanh* nonlinearity. The number of layers and number of hidden nodes per layer are selected to be 8 and 1024, re-

Table 1. WERs of the proposed R-SMAC features in contrast to other existing features for children's speech test set under varying additive noise conditions. The performance evaluation is done separately on two ASR systems, one developed using only adults' speech training data from WSJCAM0 and the other when children's speech training data from PFSTAR is also pooled in.

Training Speech	Noise Type	SNR (dB)	WER in %		
			MFCC	SMAC	R-SMAC
Adult	Clean		24.25	24.19	**22.58**
	White	5	88.65	79.41	**73.68**
		10	75.85	70.13	**65.92**
	Babble	5	93.74	90.22	**85.31**
		10	74.90	70.02	**67.30**
Adult + Child	Clean		14.63	13.47	**12.32**
	White	5	49.75	45.66	**43.34**
		10	32.73	29.53	**26.63**
	Babble	5	59.19	51.60	**47.84**
		10	37.27	32.53	**29.93**

spectively. The soft-max function is used as the output layer. An initial learning rate of 0.015 is selected which is reduced to 0.002 in 20 epochs. The minibatch size for neural net training is selected to be 512. The 40-dimensional fMLLR-normalized features are further spliced in time with context size of 9 prior to learning the DNN parameters. For evaluating the recognition performances, the word error rate (WER) metric is used. While decoding the children's test set, a 1.5k bigram language model (LM), trained on the transcripts of the speech data in PF-STAR excluding the test set is employed. The employed LM has an out of vocabulary (OOV) rate of 1.20% and perplexity of 95.8 for the children's test set, respectively. Further, a lexicon of $1,969$ words including the pronunciation variations is employed. For decoding adults' test set, the standard MIT-Lincoln 5k Wall Street Journal bigram LM is used. This LM has a perplexity of 95.3 for the adults' test set while there are no OOV words.

3.2. Results and discussions

The WERs for the children's test set on the ASR system developed using the mix of adults' and children's speech under clean test conditions are given in Table 1. For the sake of contrast, the WERs for the case when only adults' speech is used for learning the model parameters are also given in Table 1. It is evident that the R-SMAC features have resulted in improved WERs than the other two existing features explored. The WERs for the adults' test set on the two kinds of ASR systems employing MFCC features happen to be 6.20% and 12.93%, respectively. For the proposed features, on the other hand, the respective WERs turn out to be 6.30% and 12.54%. It is to note that the proposed adaptive-liftering does not result in any noticeable degradation in the recognition performances. Moreover, the pooling of children's speech data into training leads to a degradation in the recognition performance for the adults' test set.

Table 2. WERs depicting the effect of including speech enhancement module prior to the computation of acoustic feature vectors.

Noise Type	SNR (dB)	Noisy		Enhanced	
		MFCC	R-SMAC	MFCC	R-SMAC
White	5	49.75	**43.34**	42.61	**40.67**
	10	32.73	**26.63**	29.37	**25.93**
Babble	5	59.19	**47.84**	55.14	**45.24**
	10	37.27	**29.93**	34.23	**29.19**

To further validate our claims, noise robustness of the existing as well as the proposed acoustic features is also studied in this paper. Two different noises, viz. babble noise and white noise extracted from NOISEX-92 [22], were added to the test data under varying levels. The noisy test sets were then decoded on the acoustic models trained on clean speech. The WERs for this study in the case of children's mismatched testing, for two signal-to-noise (SNR) values, are also given in Table 1. It is to note that, the use SMAC features is found to be more robust to additive noise than that of the MFCC features. Furthermore, the proposed R-SMAC features are found to be superior to the other two features due to enhanced pitch robustness while retaining the immunity to noise. This observation is consistent across the two kinds of ASR systems developed in this work.

The inclusion of foreground speech segmentation and enhancement module leads to further reduction in WERs for both MFCC and proposed features. The WERs for this study with respect to the ASR system trained on the mix of adults' and children's speech are given in Table 2. Large reductions in WERs observed in the case of MFCC features than those for the R-SMAC features are attributed to the inherent noise robustness of the later ones. Yet the observed changes are significant for the low SNR cases. At the same time, the inclusion of speech enhancement module does not degrade the recognition performance for the high SNR cases.

4. CONCLUSION

This work explores the combination of a number of front-end signal processing techniques towards achieving robust recognition of children's speech. The proposed acoustic features are observed to enhance the pitch-robustness of the existing SMAC features. At the same time, the immunity towards additive noises is largely retained. To further improve the noise robustness, an earlier developed foreground speech segmentation and enhancement approach is also incorporated prior to feature extraction. The effectiveness of the proposed front-end speech parameterization approach has been verified on an ASR system developed using DNN-HMM-based acoustic modeling technique.

5. REFERENCES

[1] 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.

[2] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust ASR," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 551–554, June 2010.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Shweta Ghai and Rohit Sinha, "Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition," in *Proc. Signal Processing and Communications (SPCOM)*, 2010.

[8] Shweta Ghai and Rohit Sinha, "Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2010, pp. 7:1–7:15, January 2010.

[9] S Shahnawazuddin and Rohit Sinha, "Low-memory fast on-line adaptation for acoustically mismatched children's speech recognition," in *Proc. INTERSPEECH*, 2015.

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

[11] K. T. Deepak and S. R. M. Prasanna, "Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1204–1218, July 2016.

[12] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, August 2008.

[13] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[14] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. Elsevier, 1995.

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

[16] 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.

[17] 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.

[18] 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.

[19] Angeliki Metallinou and Jian Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners," in *Proc. INTERSPEECH*, 2014, pp. 1468–1472.

[20] 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.

[21] Kaldi Toolkit: http://kaldi.sourceforge.net.

[22] "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.