SUB-BAND SPECTRAL VARIANCE FEATURE FOR NOISE ROBUST ASR

HariKrishna Maganti, Silvia Zanon, Marco Matassoni and Alessio Brutti

Fondazione Bruno Kessler - Center for Information Technology - IRST
via Sommarive 18, 38123 Povo, Trento, Italy
{maganti, silzanon, matasso, brutti}@fbk.eu

ABSTRACT

The goal of this work is to improve automatic speech recognition (ASR) performance in very noisy and reverberant environments. The solution is based on extracting sub-band spectral variance normalization based features, which are capable of estimating the relative strengths of speech and noise components both in presence and absence of speech. The advanced ETSI-2 frontend, RASTA-PLP, MFCC alone and in combination with spectral subtraction are tested for comparison purposes. Speech recognition evaluations are performed on the noisy standard AURORA-2 and meeting recorder digit (MRD) subset of AURORA-5 databases, which represent additive noise and reverberant acoustic conditions. The results reveal that the proposed method is robust and reliable for both low SNR and reverberant scenarios, and provide considerable improvements with respect to the traditional feature extraction techniques.

1. INTRODUCTION

In practical environments, a variety of signal variabilities decrease the speech intelligibility as well as the performance of ASR systems. The important cause of signal variabilities is due to the additive noise and reverberation. Additive noises from interfering noise sources, and convolutive noise arising from acoustic environment majorly contribute to a degradation of performance in speech recognition systems.

Noise reduction techniques are used to suppress the noise and improve the perceptual quality and intelligibility of the speech. To deal with additive noise, different techniques have been proposed based on voice activity detection based noise estimation, minimum statistics noise estimation, histogram and quantile based methods, and estimation of the posteriori and a priori signal-to-noise ratio [1, 2, 3, 4]. In [4], various approaches to speech enhancement based on noise estimation and spectral subtraction are discussed. Apart from stationary background noise, another important source of degradation is caused by reverberation produced in acoustic environment. Traditional noise reduction methods, which are based on statistical modeling properties of noise and speech such as spectral subtraction, Wiener filtering, and Bayesian estimation fail to reduce reverberation effect as both clean speech and reverberated speech posses similar statistical properties. The speech signal acquired in a reverberant room can be modeled as convolution of the speech signal with the room impulse response, and several methods have been proposed to deal with convolution distortion in [5, 6, 7, 8].

Additive noise reduction techniques usually have a trade-off between the amount of noise removal and speech distortions introduced due to processing of the speech signal. The intensity of distortion induced is particularly high at very low signal-to-noise ratios (SNR), degrading the performance of ASR system. In the context of convolutive noise, dereverberated speech can be obtained by inverse filtering the impulse response of the room. However, the impulse response depends upon the distance between the speaker and the microphone, and room conditions. Thus extracting common set of robust features which can perform well at low SNRs and also handle various room impulse responses is a complex and challenging task.

The traditional Mel-frequency cepstral coefficients (MFCC) features are sensitive to noise as they capture the absolute energy response of the spectrum. To overcome this limitation, multi-resolution spectral entropy based features were proposed in [9] to represent the peak energy in each band as opposed to the mean values of the MFCC features. These features in combination with perceptual linear prediction features (PLP) improved the performance over the PLP alone, specifically at low SNRs. Variance adapted acoustic models have been used for improving the robustness of recognizers [10, 11]. In [11], variance adapted projection measure improved the ASR performance significantly at low SNRs of $10dB$, $5dB$ and $0dB$. Also, spectral subtraction has been commonly used for noise suppression for enhancing speech signals with static background noise [12, 13]. However, it has not been very successful in signal processing applications, because of the well-known musical noise effect. Musical noise even at low levels significantly affect human speech perception. Various methods have been proposed to reduce musical noise, but improvement of performance at low SNR values is not guaranteed [14]. The primary reason is because musical noise is produced by physical inconsistencies of subtracted amplitude spectra and unmodified phase. To overcome this problem, a filtering process which can attenuate or enhance spectra without separating the amplitude and phase components is required.

In this work, an alternate approach for feature extraction based on normalized variances of the speech magnitude spectrum in Mel sub-bands is used to overcome the limitation of the spectral subtraction technique. The variance is a statistical measure which represents dynamically changing spectral information, and is effective for distinguishing the speech and noise frames, thereby improving the noise robustness. This is due to the fact that, the dynamic range of the noise is smaller than the speech, because of the intrinsic nature of the speech signal. The derived features are efficient at very low SNRs and also in the context of mismatch between training and testing reverberant environments. The competency of proposed features is demonstrated with the experiments performed on all test sets of $5dB$ and $0dB$ SNRs of Aurora-2, which represent additive noise at low SNRs and real-time reverberant speech acquired through four different microphones of Aurora-5 database. The recognition results

obtained using standard advanced ETSI front-end [16] and RASTA-PLP, MFCC and combination of spectral subtraction with MFCC features are tested for comparison purposes.

The paper is presented as follows: Section 2 describes the feature extraction methodology. Section 3 presents database description and experiments and section 4 discusses the results. Finally, Section 5 concludes the paper.

2. FEATURE EXTRACTION

The proposed approach is aimed to have consistent performance in all practical environments i.e. considering both additive and convolutive noises. Noise is due to peaks and valleys which is caused by high variance in the spectrum of noise, in which the values of each frequency vary in a random way. After spectral subtraction the peaks of noise are shifted down, while valleys between two peaks are minimized. So, the peaks of noise still remain in the enhanced signal which cause musical noise. For this reason, to reduce musical noise it is necessary to reduce the peaks in the enhanced signal. For the same, in the original noisy speech signal, the peaks are suppressed based on the computation of frame variance, and by filling-in the valleys between the two peaks. This algorithm tries to process the noise using the information provided by the variance of noise frames and speech frames. In this way, the goal is to suppress the peaks of noise to reduce the musical noise and at the same time, to distinguish between clean and noisy frames of the signal. In this proposed work, the idea is to retain high variance frames and minimize low variance frames so as to emphasize the speech information in the noisy signal. The result is a better segmentation between noise frames and speech frames and a reduction of noise in the entire spectrum due to the reduction of noise peaks. The block diagram of the algorithm is shown in Fig. 1, and is as follows:

After applying spectral subtraction, as described in [15], the power spectrum is mapped onto an auditory frequency axis, by combining FFT bins into equally-spaced intervals on the Mel axis defined by:

$$\text{Mel}(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (1)$$

where f is the frequency in linear domain. The output is a Mel-scaled vector consisting of $Y_k(m)$, k=1,..,K, m=0,1,..,∞, where k is the subband number, and m is the time index of each subband signal. Then, the variance for each frame is computed as

$$v(m) = \frac{1}{K-1} \sum_{i=1}^{K} (v_i - \hat{v})^2 \quad (2)$$

where K is the number of bands, $\hat{v}$ is the mean, and v_i is element number i. To suppress the peaks of noise, these values are normalized with respect to the maximum value for complete utterance as

$$w(m) = \frac{v(m)}{\max\{v(m)\}} \quad (3)$$

Later, these normalized values are used as weights which are then multiplied with the filter bank energies as shown

$$\tilde{Y}_k(m) = Y_k(m) \cdot w(m) \quad (4)$$

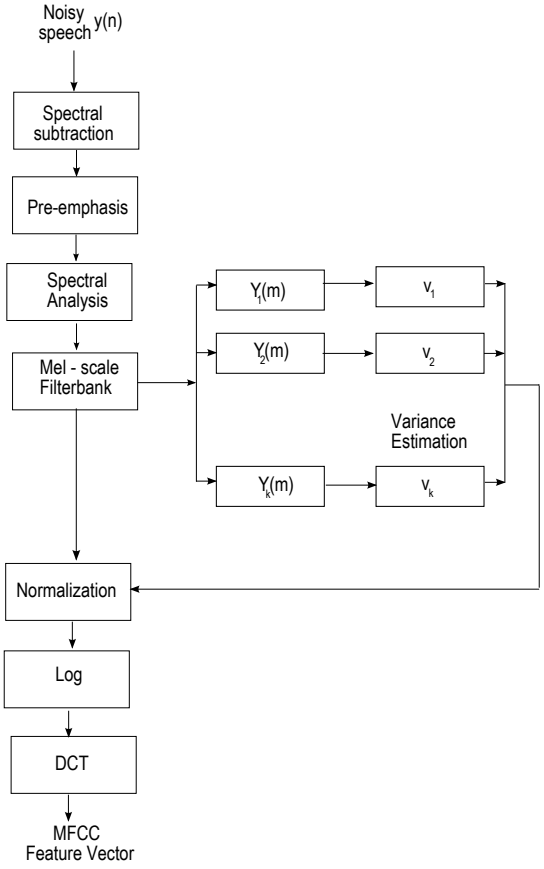

Figure 1: Processing stages of the subband variance spectral feature.

The signal is then decorrelated by applying a discrete cosine transform (DCT) for feature extraction.

For a given signal sampled at 8 kHz, short segments of speech are extracted with a 25 ms rectangular window and the window is shifted by 10 ms. Each speech frame is then processed by a 32-channel Mel filterbank. Variance calculation across all the frames and normalization are performed to derive the weights which are then multiplied with the original noisy signal. The 32 Mel spectral values are transformed to the cepstral domain by means of a DCT. Thirteen cepstral coefficients $C0$ to $C12$ are derived. $C0$ is replaced by logarithm of the energy computed from the speech samples.

Fig. 2 (a, b, c) shows waveforms, spectrograms, variance contour and sub-band spectral variance processed spectrograms of the clean speech and speech corrupted with $5dB$ and $0dB$ SNRs subway corrupted speech from Aurora2 database for an utterance "three eight eight". From second row, we can clearly observe that the unprocessed spectrogram are affected by the noise. The third row corresponds to the variance contours which is effective in providing segmentation between noise frames and speech frames and also helpful in reducing the noise in the entire spectrum due to the reduction of noise peaks. The fourth row shows figures of the spectrograms of clean and sub-band spectral variance processed speech. It can be observed that the proposed sub-band spectral variance technique is effective in reducing the noise in the spectrum and retain important information required for speech recognition.

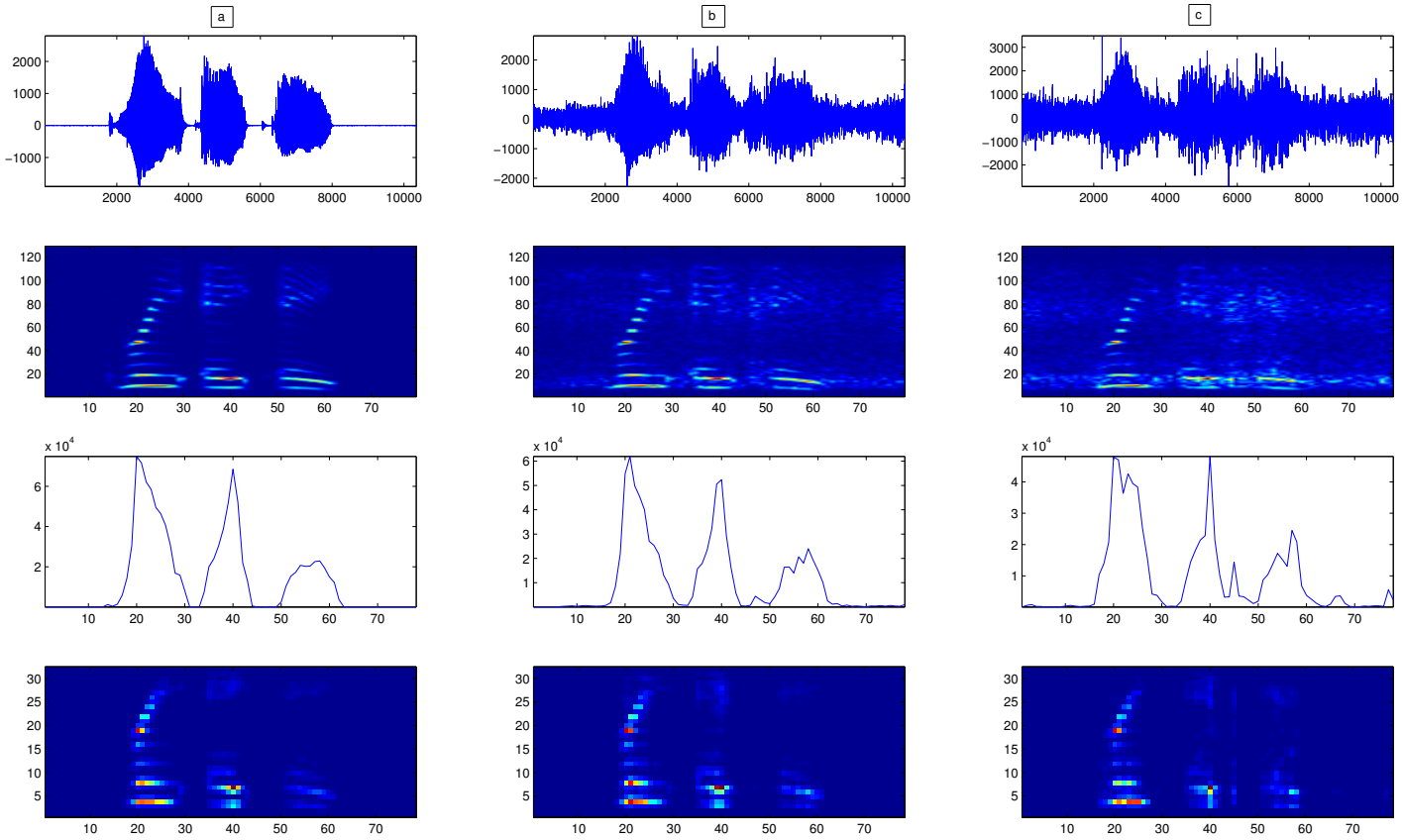

Figure 2: *Waveform, spectrogram, variance contour and sub-band spectral variance processed spectrogram of (a) clean speech (b) 5dB subway noise corrupted speech and (c) 0 dB subway corrupted speech for an utterance "three eight eight".*

The normalized variance weighting method is a smoothing technique which improves automatic speech recognition in noisy conditions where speech frames are characterized by high variance and noise frames by low variance values. The spectral subtraction method is used to reduce the broadband noise due to peaks, and the variance weighting technique is effective in reducing the musical noise by reducing the dynamic range of its magnitude spectrum, which results in improved speech recognition performance.

3. EXPERIMENTS AND RESULTS

Two sets of HMM-based ASR experiments are performed with the proposed approach using a full HTK based recognition system [17]. The first is a connected digit recognition task using the Aurora 2 database [18]. The second is a meeting recorder digits of Aurora-5 corpus [19].

3.1 Additive Noise

For the Aurora 2 experiments, training and testing follow the specifications described in [18]. A word-based ASR system for digit string recognition where each HMM word model has 16 emitting states is adopted. A three-state silence model and a one state short pause model are used. Testing data include eight types of realistic background noise subway, babble, car, exhibition hall, restaurant, street, airport and train station noise at various SNRs (clean, 20, 15, 10, 5, 0, and 5 dB). There are three test sets. Set A contains 4004 utterances in the first four types of noise, set B contains 4004 utterances in the other four, and set C contains 2002 utterances where only subway and street noise are present. The multicondition training, where 8440 utterances are split into 20 subset with 422 utterances in each subset is considered. The 20 subsets represent 4 different noise scenarios at 5 different SNRs (from 20 dB to 0 dB). To study the efficiency of the proposed features, the most challenging scenarios of low SNRs of 5*dB* and 0*dB* SNRs are tested.

	Subway		Babble		Car		Exhibition	
	0dB	5dB	0dB	5dB	0dB	5dB	0dB	5dB
ETSI-2	33.1	11.7	37.8	**12.4**	46.6	12.2	35.6	**12.4**
RASTA-PLP	32.5	**10.8**	43.2	14.7	48.1	11.6	34.9	**12.4**
MFCC	40.1	19.5	41.4	16.5	41.4	18.6	41.2	17.6
SS+MFCC	49.4	23.4	43.4	17.7	50.0	20.6	53.7	26.0
SVF	**27.5**	12.9	**37.3**	15.1	**21.5**	**10.4**	**27.8**	14.6

Table 1: Testset A results for Aurora 2 database (word error rate %)

	Restaurant		Street		Airport		Train	
	0dB	5dB	0dB	5dB	0dB	5dB	0dB	5dB
ETSI-2	40.7	16.5	38.6	14.3	34.9	13.8	43.9	16.5
RASTA-PLP	46.9	19.6	40.9	14.9	35.9	15.1	41.7	16.8
MFCC	43.3	18.1	38.3	14.4	36.3	14.5	44.9	17.1
SS+MFCC	50.3	21.9	45.4	21.9	39.2	16.6	45.4	19.5
SVF	**37.5**	**16.3**	**29.2**	**14.0**	**29.1**	**13.3**	**26.7**	**12.0**

Table 2: Testset B results for Aurora 2 database (word error rate %)

	Subway		Street	
	0dB	5dB	0dB	5dB
ETSI-2	53.2	17.6	45.6	17.4
RASTA-PLP	54.4	17.8	46.7	18.6
MFCC	58.1	20.1	48.4	19.4
SS+MFCC	59.3	28.7	53.2	27.3
SVF	**33.2**	**15.3**	**34.2**	**16.3**

Table 3: Testset C results for Aurora 2 database (word error rate %)

Tables 1,2,3 show the results in % word error rates for all testsets of the database. ETSI-2, SS+MFCC, and SVF indicate the standard ETSI advanced frontend, spectral subtraction preprocessing with MFCC, and the proposed spectral variance features. All the features considered are of standard 39-dimensions along with their delta and acceleration derivatives. The results indicated in bold represent the best performance among all the features. From Tables 1,2,3, it is evident that the classical spectral subtraction commonly used to counter additive noise has the worst performance compared to all features, indicating inefficiency and effect of musical noise at the low SNRs. It can also be seen that the traditional MFCC have better performance compared to spectral subtraction preprocessing for all the cases. The RASTA-PLP has better performance than the MFCC, indicating efficiency of these features. The advanced ETSI-2 frontend, designed specifically for Aurora-2 database has the best performance compared to all features, except the SVF features. It can also be observed from Table 1, that for majority of the cases SVF has low error rates and from Tables 2 and 3 for all the cases SVF has the low error rates. It can also be seen that for all the $0dB$ SNR cases, the SVF has low error rates. However, when tested with high SNR scenarios the performance was on par with MFCC without causing any degradation in the performance.

3.2 Convolutive Noise

The experiments are conducted on a subset of the Aurora-5 corpus - meeting recorder digits [19]. The data comprise real recordings in a meeting room, recorded in a hands-free mode at the International Computer Science Institute in Berkeley. The dataset consists of 2400 utterances from 24 speakers, with 7800 digits in total. The speech was captured with four different microphones, placed at the middle of the table in the meeting room. The recordings contain only a small amount of additive noise, providing the typical effect of hands-free recording in the reverberant room. There are four different versions of all utterances recorded with four different microphones (labeled as 6, 7, E and F), as described in [19]. The clean TI DIGITS without any additional filtering for training is considered.

	6	7	E	F
ETSI-2	36.1	53.1	42.4	37.7
RASTA-PLP	25.1	34.3	32.3	26.3
MFCC	26.3	39.7	30.0	24.4
SS+MFCC	24.0	36.0	28.7	23.7
SVF	**23.9**	**29.5**	**28.3**	**23.1**

Table 4: Word error rates (%) for different feature extraction techniques on four different microphones.

From Table 4, it is evident that the advanced ETSI front-end has the highest error rates compared to all the features. This demonstrates that for reverberant environments the advanced ETSI front-end is not effective as compared to its performance in the presence of additive background noise. It can be inferred that the techniques applicable for additive background noise removal are not suitable to handle reverberant conditions which is consistent with the studies of [20]. The combination of the spectral subtraction and MFCC has low error rates compared to MFCC and RASTA-PLP has the best performance among all the features, except the SVF features. The SVF has low error rates for all the microphones, indicating the efficiency of the proposed features in reverberant environments.

4. DISCUSSION

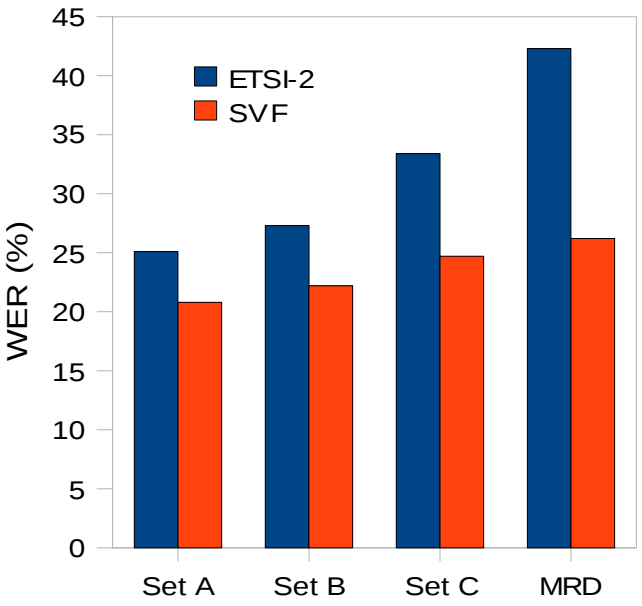

Figure 3: *Comparison of ETSI-2 and SVF features*

Fig. 3 shows comparison of the standard ETSI-2 advanced frontend and the proposed spectral variance features. The comparison is performed by considering the averages over $0dB$ and $5dB$ for all testsets of Aurora-2 database and meeting recorder digit dataset of Aurora-5 database. We can observe that the proposed spectral variance based features perform consistently better than the ETSI-2 frontend for all the testsets of Aurora-2 showing efficancy of these features

for all types of noises which include subway, babble, car, exhibition, restaurant, street, airport and train. Also, as observed from Fig. 3 the spectral variance features perform much better than the ETSI-2 advanced frontend for meeting recorder digit task, indicating efficacy of these features for reverberant scenarios. Also, this clearly demonstrates that the techniques suitable for additive background noise removal are not adequate to handle reverberant conditions. Apart from consistency and robustness, the proposed simple technique does not deteriorate the recognition performance at high SNRs nor require complex additional processing incurring additional computational costs.

5. CONCLUSIONS

In this paper, a variance based approach is proposed to improve the speech recognition performance in a noisy and reverberant environments. The proposed features were derived from subband variance normalization technique where speech frames are characterized by high variance and noise frames by low variance, which are suppressed to improve ASR performance. The spectral subtraction method was used to reduce the broadband noise due to peaks, and the variance weighing technique was effective in reducing the musical noise by reducing the dynamic range of its magnitude spectrum, which resulted in the improved speech recognition performance. The features were evaluated on Aurora-2 and Aurora-5 meeting recorder digit task. Results were compared with standard ETSI advanced front-end and conventional features. The results show that the proposed features perform consistently better both in terms of robustness and reliability.

This study raised number of issues, including on-line implementation with normalization techniques applied onto the past frames of the speech signal, improvement of variance based features to deal with both additive noise and reverberation conditions simultaneously. For the future, we like to investigate these issues to efficiently deal with real world noisy speech, and evaluate these features on large vocabulary tasks such as Aurora-4.

REFERENCES

[1] M. Woelfel and J. McDonough,"Distant Speech Recognition", John Wiley and Sons, 2009.

[2] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics", IEEE Transactions on Speech and Audio Processing, 9(5):504–512, 2001.

[3] J. Droppo and A. Acero,"Environmental Robustness", in Springer Handbook of Speech Processing, Benesty, J.; Sondhi, M. M. and Huang, Y.[Eds], 653-679, 2008.

[4] Y. Ephraim and I. Cohen, "Recent Advances in Speech Enhancement", in The Electronic Handbook, CRC Press, 2006.

[5] S. Furui, and M. Sondhi, " Advances in Speech Signal Processing", in Marcel Dekker, Inc., New York, 1991.

[6] T. Takiguchi, S. Nakamura, and K. Shikano, "Hands-free Speech Recognition by HMM Composition in Noisy Reverberant Environments", in IEICE Transactions D-II, J79-D-II (12), 2047-2053, 1996.

[7] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction Steered by Multi-step Forward Linear Prediction for Single Channel Speech Dereverberation", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), I, 817-820, 2006.

[8] H. K. Maganti, M. Matassoni, "An Auditory based Modulation Spectral Feature for Reverberant Speech Recognition", in Proceedings INTERSPEECH-2010, 570-573.

[9] H. Misra, S. Ikbal, S. Sivadas and H. Bourlard, "Multi-resolution Spectral Entropy Feature for Robust ASR", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), I, 253-256, 2005.

[10] M. Delcroix, T. Nakatani and S. Watanabe,"Static and Dynamic Variance Compensation for Recognition of Reverberant Speech With Dereverberation Preprocessing", in Transactions of Audio, Speech, and Language Processing, II(17), 324-334, 2009.

[11] C. Jen-Tzung, L. Lee-Min and W. Hsiao-Chuan, "Noisy Speech Recognition using Variance Adapted Likelihood Measure", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), I, 45-48, 1996.

[12] S. F. Boll, Suppression of Acoustic Noise in Speech using Spectral Subtraction, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP 27: 113120, April 1979.

[13] K. Paliwal, K. Wojcicki and B. Schwerin, "Single-channel Speech Enhancement using Spectral Subtraction in the Short-time Modulation Domain" in Speech Communication, vol. 52, May 2010.

[14] L. Arslan, A. McCree, and V. Viswanathan, "New Methods for Adaptive Noise Suppression", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Detroit, May 1995.

[15] M. Berouti, R. Schwartz and J. Makhaul, " Enhancement of Speech Corrupted by Acoustic Noise", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 208-211, 1979.

[16] ETSI ES 202050, "STQ; Distributed Speech Recognition, Advanced Front-End Feature Extraction Algorithm, Compression Algorithm", ETSI ES 202 050 v1.1.3 (2003-11), Nov. 2003.

[17] S. J. Young and et al "The HTK Book (version 3.4)", Cambridge University Engineering Department, 2007.

[18] H. G. Hirsch and D. Pearce, "Applying the Advanced ETSI frontend to the Aurora-2 task" in version 1.1, 2006.

[19] H. G. Hirsch, "Aurora-5 Experimental Framework for the Performance Evaluation of Speech Recognition in Case of a Hands-free Speech Input in Noisy Environments" Online: http://aurora.hsnr.de/aurora-5/reports.html.

[20] H. G. Hirsch and H. Finster, "A New Approach for the Adaptation of HMMs to Reverberation and Background Noise" in Speech Communication, vol. 50, March 2008.