Noise Spectrum Estimation using Gaussian Mixture Model-based Speech Presence Probability for Robust Speech Recognition

M. J. Alam[1,2], *P. Kenny*[1], *P. Dumouchel*[2], *D. O'Shaughnessy*[3]

[1] CRIM, Montreal, Canada
[2] ETS, Montreal, Canada
[3]INRS-EMT, Montreal, Canada

{jahangir.alam, patrick.kenny}@crim.ca, Pierre.Dumouchel@etsmtl.ca, dougo@emt.inrs.ca

10.21437/Interspeech.2014-162

Abstract

This work presents a noise spectrum estimator based on the Gaussian mixture model (GMM)-based speech presence probability (SPP) for robust speech recognition. Estimated noise spectrum is then used to compute a subband *a posteriori* signal-to-noise ratio (SNR). A sigmoid shape weighting rule is formed based on this subband *a posteriori* SNR to enhance the speech spectrum in the auditory domain, which is used in the Mel-frequency cepstral coefficient (MFCC) framework for robust feature, denoted here as Robust MFCC (RMFCC) extraction. The performance of the GMM-SPP noise spectrum estimator-based RMFCC feature extractor is evaluated in the context of speech recognition on the AURORA-4 continuous speech recognition task. For comparison we incorporate six existing noise estimation methods into this auditory domain spectrum enhancement framework. The ETSI advanced front-end (ETSI-AFE), power normalized cepstral coefficients (PNCC), and robust compressive gammachirp cepstral coefficients (RCGCC) are also considered for comparison purposes. Experimental speech recognition results show that, in terms of word accuracy, RMFCC provides an average relative improvements of 8.1%, 6.9% and 6.6% over RCGCC, ETSI-AFE, and PNCC, respectively. With GMM-SPP -based noise estimation method an average relative improvement of 3.6% is obtained over other six noise estimation methods in terms of word recognition accuracy.

Index Terms: speech recognition, speech presence probability, noise spectrum estimation, GMM, IMCRA

1. Introduction

The objective of noise robust speech recognition is to maintain satisfactory recognition performance when the test environment is different from the training environment, i.e., under mismatch train/test conditions. The mismatched conditions are due to corruption of speech signals by acoustic background noise, channel frequency response, different channel, and acoustic reverberation. Various research in the literature has been done to improve the robustness of speech recognition systems under mismatched conditions. The methods to compensate for the effects of environmental mismatch can be implemented at the front-end (feature extractor) or at the back-end or both. Robust features can be obtained by appending a pre-processing step, like speech enhancement [2-4], or by incorporating noise compensation algorithms [5-7] in a conventional mel-frequency cepstral coefficients (MFCC) [1] framework. Estimation of noise spectrum from the corrupted speech signal is a fundamental component of many robust speech processing applications, including the speech enhancement, robust front-ends for speaker and speech recognition systems. Under adverse acoustic environments involving low signal-to-noise ratio (SNR) conditions, non-stationary noise environments (e.g., babble noise), and weak speech components the robustness of speech enhancement and feature extractors for speech recognition systems is significantly affected by the capability to reliably and accurately estimate the noise spectrum.

Several methods have been introduced in the literature for the estimation of noise spectrum. The simplest approach to noise spectrum estimation is to average the noisy speech spectrum over non-speech segment using a voice activity detector. The minimum statistics (MS) noise estimator [8] estimates the noise spectrum as the minima values of a smoothed spectrum estimate of the noisy speech signal within a finite window. In [9], a computationally efficient minima tracking scheme is proposed which is slow in updating the noise estimate especially when there is a sudden change in noise level. The minima controlled recursive averaging (MCRA) [10] obtains the noise spectrum estimate by averaging past spectral values using a smoothing parameter adjusted by the probability of speech presence. Compare to MCRA the improved MCRA (IMCRA), proposed in [12], utilizes time-varying smoothing parameter that depends on an estimate of the speech presence probability (SPP). Another improved MCRA (IMCRA2) is presented in [13] based on a conditional MAP (maximum *a posteriori*) criterion by taking full consideration of the inter-frame correlation of voice activity. A modified version of MMSE (minimum mean square error) -based noise estimator is proposed in [14] that uses a soft SPP with fixed priors.

In this work we use a SPP-based noise spectrum estimation method for robust features extraction for speech recognition. Speech presence probability for each frequency bin is estimated using a sequential Gaussian Mixture Model (SGMM) [15] in an unsupervised learning framework. Estimated noise spectrum is utilized to form a sigmoid shaped suppression rule based on a subband *a posteriori* SNR to enhance the speech spectrum in the auditory domain [6, 16]. This auditory domain spectrum enhancement (ASE) technique is incorporated in a MFCC feature extraction framework to extract robust feature, dubbed as robust MFCC (RMFCC), for speech recognition. The recognition performance of the RMFCC is compared with that of the PNCC [5], ETSI-AFE [7] and RCGCC [16] front-ends. In order to show the effectiveness of the GMM-SPP -based noise estimation method, in the context of speech recognition performance on the AURORA-4 corpus [17, 18], the following six noise estimators are considered: IMCRA [12], IMCRA2 [13], MS [8], MMSE-SPP [14], Doblinger's subbnad minima tracking [9], and the connected frequency regions-based noise estimator [13]

2. Speech Presence Probability-based Noise Spectrum Estimation

In order to estimate the speech presence probabilities we use a sequential GMM-based unsupervised learning framework proposed in [15] for speech/nonspeech discrimination. Let Y_m represent the log power spectrum for the m-th frame of a DFT (discrete Fourier Transform) subband, z is the nonspeech/speech label, $z \in \{0,1\}$, where 0 for nonspeech and 1 for speech. The log power spectrum is smoothed using a median filter with window of 5-frames. Gaussian Mixture Model used comprised two Gaussian distributions, each trying to model either nonspeech or speech. The models were trained using an unsupervised learning process [15], whereby the initial frames (usually first sixty frames, if the number of frames of an utterance is less than sixty then half of the total frames is taken as initial frames) from a signal were clustered into the two Gaussians, with the distribution with the lowest mean representing nonspeech regions and the distribution with the higher mean representing speech regions. The estimated distributions were also used to determine a decision threshold to discriminate speech from non-speech. Usually, it is chosen as the point between two centers where the probabilities are equal.

The likelihood of Y_m given speech/nonspeech GMM model $\lambda_m = \{w_{m,z}, \mu_{m,z}, \Sigma_{m,z}\}$ $p(Y_m|z,\lambda_m)$ is given by

$$p(y_m|z,\lambda_m) = \frac{1}{\sqrt{2\pi\Sigma_{m,z}}} \exp\left\{-\frac{(Y_m-\mu_{m,z})^2}{2\Sigma_{m,z}}\right\} \quad (1)$$

where $w_{m,z}$, $\mu_{m,z}$, and $\Sigma_{m,z}$ denote the mixture weight, mean and variance, respectively. The initial GMMs are firstly established using the typical EM (expectation maximization) algorithm, and then sequentially estimated at each frame using the following formulae [15]:

$$w_{m+1,z} = \alpha w_{m,z} + (1-\alpha)\, p(z|Y_{m+1},\lambda_m), \quad (2)$$

$$\mu_{m+1,z} = \frac{\alpha w_{m,z}\mu_{m,z} + (1-\alpha)\, p(z|Y_{m+1},\lambda_m) Y_{m+1}}{w_{m+1,z}}, \quad (3)$$

$$\Sigma_{m+1,z} = \frac{\alpha w_{m,z}\Sigma_{m,z} + (1-\alpha)\, p(z|Y_{m+1},\lambda_m)(Y_{m+1}-\mu_{m+1,z})}{w_{m+1,z}}, \quad (4)$$

$$p(z|Y_m,\lambda_m) = \frac{w_{m,z}\, p(Y_m|z,\lambda_m)}{\sum_z w_{m,z}\, p(Y_m|z,\lambda_m)}, \quad (5)$$

where $p(z=1|Y_m,\lambda_m)$ denotes the speech presence probability (SPP) for the m-th frame.

If $P(k,m)$ represents the estimated SPP for the k-th frequency bin index and m-th frame then an estimate for the noise power spectrum $\hat{D}(k,m)$ is given by the following recursive averaging

$$\hat{D}(k,m) = \eta \hat{D}(k,m-1) + (1-\eta)\hat{D}_1(k,m), \quad (6)$$

where η is the smoothing parameter and $\hat{D}_1(k,m)$ is expressed as:

$$\hat{D}_1(k,m) = P(k,m)\hat{D}(k,m-1) + (1-P(k,m))Y(k,m), \quad (7)$$

where $Y(k,m)$ is the noisy speech power spectrum for the m-th frame and k-th frequency index. In this work we use $\eta = 0.8$. In this method, an initial estimate of the noise power spectrum is computed by averaging the first ten frames of the observed speech spectrum. Fig. 1 shows the estimated SPP at the frequency bin index k = 10 (around 281 Hz) for (i) a clean speech signal and (ii) a noisy speech signal degraded with car noise with SNR = 5 dB.

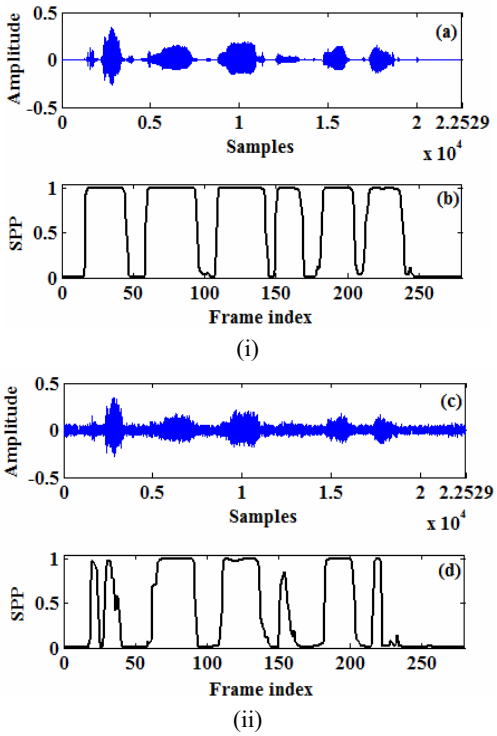

Figure 1: Speech presence probabilities at a frequency bin index k = 10 (around 281 Hz) obtained using sequential GMM (Gaussian mixture model) for (i) clean speech signal, and (ii) a noisy speech signal corrupted with car noise at a signal-to-noise ratio of 5 dB.

Fig. 2 presents a complete block diagram of the robust MFCC (RMFCC) front-end that use GMM-SPP noise estimator for the estimation of noise power spectrum. Based on the noisy speech power spectrum and the estimated noise power spectrum a sigmoid shape weighting rule is formed and can be expressed as:

$$W(n,m) = \frac{1}{1+\exp\left\{-\frac{(\gamma_{sb}(n,m)-c)}{a}\right\}}, \quad (8)$$

where n is the mel-subband index, m is the frame index, $1/a$ and c are the slope and mean, respectively. $\gamma_{sb}(n,m)$ is the subband *a posteriori* SNR given by:

$$\gamma_{sb}(n,m) = \max\left(10\log_{10}\left(\frac{Y_{as}(n,m)}{D_{as}(n,m)}\right), -4.0\right),$$

where $D_{as}(n,m)$ and $Y_{as}(n,m)$ are the Mel-scale triangular shaper filterbank integrated noise and noisy speech auditory spectrum. Here, $a = c = 4.5$ is chosen experimentally [6, 16].

The enhanced auditory spectrum can then be obtained by multiplying (element-wise) $Y_{as}(n,m)$ with $W(n,m)$. Similar to [5, 6, 16], a power function nonlinearity with a coefficient of 1/15 is then applied to approximate the relationship between a human's perception of loudness and the sound intensity. For feature normalization a short-time mean and scale normalization (STMSN) [16, 19] technique is used with a sliding window of 1.5 seconds. It, under mismatched conditions, helps to remove the difference of log spectrum between the training and test environments by adjusting the short-time mean and scale.

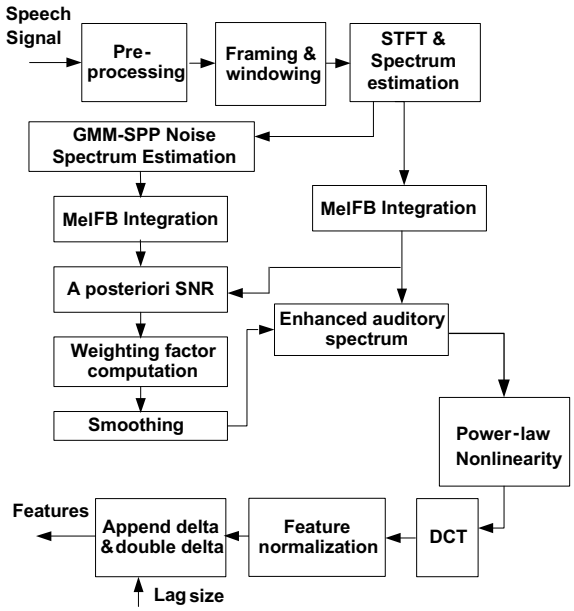

Figure 2: Block diagram of the Robust Mel-Frequency Cepstral Coefficients (RMFCC) feature extractor incorporating GMM-SPP (Gaussian mixture model-speech presence probability) noise spectrum estimation method for auditory spectrum enhancement (ASE).

3. Baseline Noise Estimation Methods

To evaluate and compare the performance of the GMM-SPP noise spectrum estimator in terms of the speech recognition accuracy six existing noise estimation methods are chosen as baselines and briefly described below.

3.1. MMSE-SPP [14]

In [14] a modified version of the MMSE-based noise estimator is proposed in which the hard decision of the voice activity detector (VAD) is replaced by a soft SPP with fixed priors. The advantage of this method is that it does not require a bias correction term as required by a MMSE-based noise spectrum estimation method; it also results in less overestimation of noise power and is computationally less expensive.

Now, replacing the GMM-SPP noise estimation method with a MMSE-SPP noise estimator in fig. 2 we extract RMFCC features and denote this as the ***RMFCC 1*** features.

3.2. IMCRA [12]

The IMCRA method, proposed in [12] is an improved version of the minima controlled recursive averaging (MCRA) noise estimator introduced in [10]. The estimate of the SPP in IMCRA is controlled by the minima values of a smoothed power spectrum of the noisy speech signal. Apart from exponential smoothing in the time direction, some averaging over neighboring frequency bins is performed, taking into account the strong correlation of speech presence in neighboring frequency bins of consecutive frames [12]. A fixed bias compensation factor is used for the minima of the smoothed spectrum. IMCRA uses two iterations of smoothing and minimum tracking, in order to make the minimum tracking during speech activity more robust [21].
The RMFCC features obtained using IMCRA noise estimator is denoted here as the ***RMFCC 2***.

3.3. IMCRA2 [13]

The IMCRA2, introduced in [13], is another improved version of the MCRA method based on a conditional MAP (maximum *a posteriori*) criterion by taking full consideration of the inter-frame correlation of voice activity. In this method noise power spectrum is obtained by the recursive smoothing parameter using SPP conditioned on both the current observation and speech activity decision in the previous frame [13].

The RMFCC features extracted by incorporating the IMCRA2 noise estimator in place of GMM-SPP is denoted here as the ***RMFCC 3***.

3.4. Minimum Statistics (MS) [8]

The MS method [8] uses minima of the smoothed periodogram of the noisy speech to estimate the noise level for each frequency bin within a finite duration window and then the result is multiplied by a factor that compensates the bias. A typical size of the window is in the order of 1 s. This technique was originally motivated by the observation in which speech and noise are usually statistically independent and the power of a noisy signal frequently decays to the power level of noise.

When the MS technique is used to estimate the noise power spectrum in the RMFCC feature extraction framework then we obtain ***RMFCC 4*** features.

3.5. Minima Tracking [9]

In [9], a computationally efficient minima tracking scheme is proposed which is slow in updating the noise estimate especially when there is a sudden change in noise level.

We denote the RMFCC features obtained with this noise estimation method as the ***RMFCC 5***.

3.6. Connected Frequency Region [11]

In this noise estimation method, the connected time-frequency regions of speech presence are used to achieve noise periodogram estimates in the regions where speech is absent. In the remaining regions, where speech is present, minimum tracks of the smoothed noisy speech periodogram are bias compensated with a factor that is updated in regions with speech absence [11].

Incorporation of this noise estimation method yields the ***RMFCC 6*** features.

4. Performance Evaluation

To evaluate speech recognition performance of the GMM-SPP -based RMFCC front-end and to compare its performance experiments were carried out on the AURORA-4 LVCSR

corpus [17, 18] using the clean training condition. Results are reported on the four different evaluation conditions mentioned in [17, 18]. Word error rate (WER) is used as an evaluation metric.

4.1. The AURORA-4 corpus

The AURORA-4 continuous speech recognition corpus, derived from the Wall Street Journal (WSJ0) corpus, has 14 test sets grouped into the following 4 groups [17, 18]: (a) Test set A - clean speech in training and clean speech in test, same channel (set 1), (b) Test set B - clean speech in training and noisy speech in test, same channel (sets 2-7), (c) Test set C - clean speech in training and clean speech in test, different channel (set 8), and (d) Test set D - clean speech in training and noisy speech in test, different channel (sets 9-14). The number inside the brackets represents the test set number defined in the AURORA-4 corpus [17, 18].

4.2. Experimental setup

For the continuous speech recognition task on the AURORA-4 corpus, all experiments employed state-tied crossword speaker-independent triphone acoustic models with 16 Gaussian mixtures per state. A single-pass Viterbi beam search-based decoder was used along with a standard 5K lexicon and bigram language model with a prune width of 250 [16, 23]. The HTK [22]-based recognizer is used for training and decoding tasks. For our experiments, we use 13 static features (including the 0th cepstral coefficient) augmented with their delta and double delta coefficients, making 39-dimensional feature vectors. The analysis frame length is 25 ms with a frame shift of 10 ms. The delta and double features were calculated using a 5-frame window. For all front-ends only static features were normalized by the feature normalization method and then dynamic features were computed from them. For the PNCC [5] and ETSI-AFE [7] front-end features were extracted following the same procedure as mentioned in the respective references.

4.3. Results and Discussion

At first we evaluate and compare the performance of the RMFCC front-end (GMM-SPP noise estimation-based robust MFCC features) with the conventional MFCC [1], PLP [23], RMCC (regularized MVDR spectrum-based cepstral coefficients) [20], PNCC [5], ETSI-AFE [7], and RCGCC (robust compressive gammachirp cepstral coefficients) [16] front-ends.

Table 1 depicts the word accuracies obtained by the RMFCC and the baseline features extractors on the AURORA-4 corpus for the four evaluation conditions. It is observed from table 1 that the RMFCC outperformed the baseline robust front-ends (i.e., PNCC, RCGCC, and ETSI-AFE) in all evaluation conditions. Among all the front-ends presented in table 1 the RMFCC performed the best in recognition accuracy. An average relative improvement in word recognition accuracy achieved by the RMFCC over the PNCC, ETSI-AFE and RCGCC features are 6.6%, 6.9% and 8.1%, respectively.

Secondly, we show the effectiveness of using GMM-SPP -based noise estimator to estimate noise power spectrum for the RMFCC feature extraction in terms of the speech recognition accuracy. In order to do that we chose six existing noise estimation approaches [8, 9, 11, 12, 13, 14] and then six versions of the RMFCC features, as mentioned in section 3, were extracted based on those noise estimation methods. Table 2 presents the word accuracies obtained by the RMFCC and the other six variants of RMFCC features (*RMFCC 1* to *RMFCC 6*). It is seen from table 2 that the GMM-SPP (i.e., RMFCC features) provided, specifically in the mismatched conditions i.e., test sets B and D and on the average, the best recognition accuracy over all baseline noise estimation methods. Average relative improvements obtained by the RMFCC over the RMFCC 1, RMFCC 2, RMFCC 3, RMFCC 4, RMFCC 5, and RMFCC 6 are 3.7%, 2.0%, 3.0%, 3.7%, 3.4%, and 5.6%, respectively. Therefore, use of the noise estimation method based on the GMM-SPP helped to improve the robustness of the RMFCC feature extractor.

Table 1. Word accuracies obtained for the robust MFCC front-end and the baseline features extractors. The higher the word accuracy the better is the performance of the front-ends.

	Word Accuracy (%)				
	A	**B**	**C**	**D**	**Avg.**
MFCC	90.02	49.19	71.12	35.44	61.44
PLP(HTK)	89.72	50.41	74.44	39.64	63.55
RMCC	**90.06**	54.25	78.23	40.63	65.79
PNCC	88.64	69.85	**81.07**	60.00	74.89
ETSI-AFE	88.59	69.58	79.52	61.51	74.80
RCGCC	88.90	68.87	80.94	59.25	74.49
RMFCC	**88.91**	**72.78**	**81.07**	**63.48**	**76.56**

Table 2: Word accuracies obtained for the various noise spectrum estimation methods-based Robust MFCC (RMFCC) front-end. The higher the word accuracy the better is the performance of the front-ends.

	Word Accuracy (%)				
	A	**B**	**C**	**D**	**Avg.**
RMFCC	88.91	**72.78**	81.07	**63.48**	**76.56**
RMFCC 1	88.73	70.85	81.10	61.95	75.66
RMFCC 2	**89.10**	71.67	80.37	63.14	76.07
RMFCC 3	88.80	71.06	**81.88**	61.58	75.83
RMFCC 4	88.95	70.80	81.33	61.54	75.65
RMFCC 5	88.88	70.99	81.58	61.50	75.74
RMFCC 6	88.58	69.99	80.85	61.25	75.17

5. Conclusions

In this paper, we presented Gaussian Mixture Model-Speech presence probability (GMM-SPP) noise estimation approach based robust MFCC (RMFCC) feature extractor. The performance of the RMFCC front-end were evaluated and compared with the robust front-ends such as ETSI-AFE, RCGCC and PNCC and six variants of RMFCC features based on the six existing noise estimators under clean training modes of the AURORA-4 LVCSR corpus. It is found that use of the GMM-SPP noise estimator helped to boost the word recognition accuracy. Experimental speech recognition results demonstrated that, in terms of word accuracy, RMFCC provides an average relative improvements of 8.1%, 6.9% and 6.6% over RCGCC, ETSI-AFE, and PNCC, respectively. With GMM-SPP -based noise estimation method an average relative improvement of 3.6% is obtained over other six noise estimation methods in terms of word recognition accuracy.

6. References

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

[2] W. Zhu, D. O’Shaughnessy, “Using noise reduction and spectral emphasis techniques to improve ASR performance in noisy conditions,” Proc. ASRU 2003, US Virgin Islands, Nov. 2003.

[3] W. Zhu, D. O’Shaughnessy, “Incorporating frequency masking filtering in a standard MFCC feature extraction algorithm,” Proc. ICSP, pp. 617-620, Beijing, Aug.-Sep., 2004.

[4] J. Droppo, A. Acero, “Environmental robustness,” in springer handbook of speech processing, Benesy, J.; Sondhi, M. M. and Huang, Y. [Eds], pp. 653-679, 2008.

[5] C. Kim and R. M. Stern., “Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring,” In IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 4574-4577, March 2010.

[6] M. J. Alam, P. Kenny, D. O'Shaughnessy, "Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum," Proc. INTERSPEECH, Portland Oregon September 2012.

[7] ETSI ES 202 050, Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; advanced front-end feature extraction algorithm; Compression algorithms; 2003.

[8] Martin, R., "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Transactions on Speech and Audio Processing, vol. 9(5), pp. 504-512, 2001.

[9] Doblinger, G., "Computationally efficient speech enhancement by spectral minima tracking in subbands," Proc. Eurospeech, 2, pp. 1513-1516, 1995.

[10] I. Cohen, B. Berdugo, "Noise Estimation by minima controlled recursive averaging for robust speech enhancement," IEEE Sig. Process. Letters, vol. 9 (1), pp. 12-15, 2002.

[11] Sorensen, K. and Andersen, S., "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions," EURASIP J. Appl. Signal Process., 18, pp. 2954-2964, 2005.

[12] Cohen, I., "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Transactions on Speech and Audio Processing, vol. 11(5), pp. 466-475, 2003.

[13] Jong-Mo Kum, Yun-Sik Park, Joon-Hyuk Chang, "Improved MCRA technique using conditional MAP criterion for speech enhancement", DSP journal, vol. 20 (6), pp. 1572-1578, Feb 2010.

[14] Timo Gerkmann, Richard C. Hendrikes, “Noise power estimation based on the probability of speech presence,” Proc. IEEE WASPAA, pp. 145-148, New York, Oct. 2011.

[15] Dongwen Ying, Yonghong Yan, Jianwu Dang, Frank K Soong, "Voice Activity Detection Based on an Unsupervised Learning Framework," IEEE Trans. on ASLP, vol. 19, no. 8, November 2011.

[16] M. J. Alam, P. Kenny, D. O'Shaughnessy, "Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique," Digital Sig. Process. Journal, March 2014.

[17] N. Parihar, J. Picone, D. Pearce, H.G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," Proceedings of the European Signal Processing Conference, Vienna, Austria, 2004.

[18] S.-K. Au Yeung, M.-H. Siu, "Improved performance of Aurora-4 using HTK and unsupervised MLLR adaptation," Proceedings of the Int. Conference on Spoken Language Processing, Jeju, Korea, 2004.

[19] Alam, J., Ouellet, P., Kenny, P., O’Shaughnessy, D., “Comparative Evaluation of Feature Normalization Techniques for Speaker Verification,” *Proc NOLISP*, LNAI 7015, pp. 246-253, Las Palmas, Spain, November 2011.

[20] Md. Jahangir Alam, Patrick Kenny, Douglas O'Shaughnessy, "Speech Recognition Using Regularized Minimum Variance Distortionless Response Spectrum Estimation-Based Cepstral Features," Proc. ICASSP, Vancouver, Canada, May, 2013.

[21] J. S. Erkelens, R. Heusdens, "Tracking of Nonstationary Noise Based on Data-Driven Recursive Noise Power Estimation," IEEE ASLP, vol. 16 (6), pp. 1112-1123, August, 2008.

[22] S. J. Young et al., HTK Book, Entropic Cambridge Research Laboratory Ltd., 3.4 edition, 2006. online: http://htk.eng.cam.ac.uk/.

[23] H. Hermansky, “Perceptual linear prediction analysis of speech,” J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738–1752, Apr. 1990.