IMPROVED AUTOCORRELATION-BASED NOISE ROBUST SPEECH RECOGNITION USING KERNEL-BASED CROSS CORRELATION AND OVERESTIMATION PARAMETERS

G. Farahani[1], S.M. Ahadi[1] and M.M. Homayounpour[2]

[1]Electrical Engineering Department, [2]Computer Engineering Department
Amirkabir University of Technology
Hafez Ave., Tehran 15914, Iran
phone: + (9821) 44491790, 64543336, 64542722, fax: + (9821) 44491790, 66406469, 66495521
email: f8023953@aut.ac.ir, sma@aut.ac.ir, homayoun@ce.aut.ac.ir

ABSTRACT

This paper proposes a new algorithm to consider cross correlation between noise and clean speech signal when autocorrelation-based features have been used for robust speech recognition. Also, an overestimation parameter has been inserted in clean speech autocorrelation estimation. We have also adopted the normalization of mean and variance of energy and cepstral parameters as an extra means of further improving the speech recognition rate.

We recently proposed a new approach for Autocorrelation-based Noise Subtraction (ANS). This method did not consider any possible cross correlation between noise and the clean speech signal. In this paper we have tried to consider this term during the estimation of clean speech signal autocorrelation.

Our results on the Aurora2 corpus have shown that the recognition rate, when the cross correlation term is considered, is improved. Furthermore, taking into account the overestimation parameter further improves the results.

1. INTRODUCTION

A main problem in the Automatic Speech Recognition (ASR) systems is the sensitivity of these systems to the environmental variations. When the speech recognition system is trained with clean speech data, its performance may degrade in real environment. The real environmental variations include the additive noise, channel distortion, voice reverberation and other interferences.

Generally, the degradation in the speech recognition rate decreases with an increase in the signal to noise ratio (SNR). In low SNRs, this performance degradation is more obvious due to the larger mismatch between the train and test conditions.

The most usual case of mismatch between the train and test conditions is the availability of additive noise. Therefore, most of the robust recognition methods assume the noise to be additive in frequency domain and stationary. In this paper, we will also consider the additive stationary noise case.

The autocorrelation domain is known as a domain with certain robustness properties in noisy conditions. Some methods extract the speech features using the spectrum extracted from the signal autocorrelation sequence. Examples of such approaches include Short-time Modified Coherence (SMC) [1], One Sided Autocorrelation LPC (OSALPC) [2] and Relative Autocorrelation Sequence (RAS) [3]. More recently, methods such as Autocorrelation Mel Frequency Cepstral Coefficient (AMFCC) [4] and Differentiation of Autocorrelation Sequence (DAS) [5] have used the amplitude of autocorrelation sequence. There also exist some methods that have tried to use the phase of autocorrelation spectrum for feature vector extraction such as Phase AutoCorrelation (PAC)[6].

Although many efforts have been made to extract robust features from the signal autocorrelation sequence, the above mentioned methods suffer from disadvantages that prevent them from achieving the best performance among other methods.

In this paper, first we will discuss the Autocorrelation-based Noise Subtraction method (ANS) [7]. This method assumes that there is no correlation between the clean speech signal and noise. We propose a modified approach by removing the uncorrelated speech and noise constraint used in ANS. Therefore, a cross correlation term is inserted in the formulation of the ANS method.

Secondly, we apply an overestimation parameter to the noise autocorrelation sequence estimation. Similar to the Spectral Subtraction (SS) method [8], here, there could exist some unwanted errors such as phase, amplitude and cross correlation [9]. The solutions we provide here, are in fact introduced to overcome such error terms. Meanwhile, unlike spectral subtraction, in autocorrelation domain we do not need the flooring parameter. This means that we do not need to be concerned about an important problem, known as one of the major limitations in spectral subtraction.

The organization of this paper is as follows. In section 2, the autocorrelation theory in relation to our overall proposed approach will be discussed. In section 3, we will describe the proposed algorithms and the parameter settings in our implementation. Section 4 includes the implementations and experiments and Section 5 concludes the paper.

2. FORMULATION OF NOISY AND CLEAN SIGNALS AND NOISE IN AUTOCORRELATION DOMAIN

2.1 Autocorrelation and ANS Method

If *v(m,n)* is the additive noise and *x(m,n)* the noise-free speech signal, then the noisy speech signal, *y(m,n),* can be written as

$$y(m,n) = x(m,n) + v(m,n) \quad 0 \le m \le M-1, 0 \le n \le N-1 \quad (1)$$

where *N* is the frame length, *n* is the discrete-time index in a frame, *m* is the frame index and *M* is the number of frames.

If *x(m,n)* and *v(m,n)* are considered uncorrelated, the autocorrelation of the noisy speech can be expressed as

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(m,k) \quad 0 \le m \le M-1, \quad 0 \le k \le N-1 \quad (2)$$

where $r_{yy}(m,k)$, $r_{xx}(m,k)$ and $r_{vv}(m,k)$ are the short-time autocorrelation sequences of the noisy speech, clean speech and noise respectively and *k* is the autocorrelation sequence index within each frame. Assuming the additive noise to be stationary, its autocorrelation sequence can be considered the same for all frames. Therefore, the frame index, *m*, can be omitted from the additive noise part in equation (2). Hence

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(k) \quad 0 \le m \le M-1, \quad 0 \le k \le N-1 \quad (3)$$

For simplicity, hereafter, we will omit the frame index *m* from the noise autocorrelation sequence.

In ANS, we assume the noise signal to be stationary. Furthermore, the average of a few initial values of the autocorrelation sequence of noisy speech signal were used for noise autocorrelation estimation, i.e..

$$\hat{r}_{vv}(k) = \frac{\sum_{i=0}^{P-1} r_{yy}(i)}{P}, \quad (4)$$

where *P* is the number of initial frames in each utterance and $\hat{r}_{vv}(k)$ is the noise autocorrelation estimation. Therefore, *ANS* estimates autocorrelation of clean speech signal as

$$\hat{r}_{xx}(m,k) = r_{yy}(m,k) - \hat{r}_{vv}(k). \quad (5)$$

2.2 Cross Correlation and Kernel Method

Generally, assuming the speech signal and noise to be completely uncorrelated, we write the autocorrelation of the noisy speech signal as the sum of the autocorrelations of clean speech signal and noise. But here, removing the above assumption, the correlation between the clean speech signal and noise should also be considered in equation (3), therefore we have:

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(k) + E\{x(m,k).v^*(m,k)\} + E\{x^*(m,k).v(m,k)\}$$
$$= r_{xx}(m,k) + r_{vv}(k) + 2.E\{|x(m,k)|.|v(m,k)|.\cos\theta(m,k)\} \quad (6)$$

where $\theta(m,k)$ is the instantaneous phase difference between clean speech signal $x(m,k)$ and noise $v(m,k)$. If we assume $\theta(m,k)$ to have a uniform distribution between $-\pi$ and π and clean speech signal and noise are stable in the averaging period, then we will have the following equation:

$$r_{yy}(m,k) \approx r_{xx}(m,k) + r_{vv}(k) + 2.|x(m,k)|.|v(m,k)|.\cos\theta(m,k) \quad (7)$$

Generally, *ANS* method assumes that there is no cross correlation between noise and clean speech signal leading to the following relationship between noisy speech, clean speech and noise autocorrelation sequences, which is the same as equation (3),

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(k). \quad (8)$$

The definition of Equation (8) is based on the assumption that the expectation value of $\cos\theta(m,k)$ in equation (7) is equal to zero. However, this might not be true even if the noise autocorrelation is estimated accurately.

It is important to consider the phase difference $\theta(m,k)$ between clean speech signal and noise in the noisy speech signal. In some situations, the autocorrelation sequence of the clean speech signal could not be exactly retrieved by ANS method, since the cross correlation between clean speech signal and noise is not taken into account. If we insert this phase difference into equation (8), we will have it as follows [11].

$$r_{xx}(m,k) = r_{yy}(m,k) - r_{vv}(k).(1 + 2r(m,k).\cos\theta(m,k))$$
$$= r_{yy}(m,k) - M(r(m,k),\theta(m,k)).r_{vv}(k) \quad (9)$$

where

$$r(m,k) = \frac{|x(m,k)|}{|v(m,k)|}$$
$$M(r(m,k),\theta(m,k)) = 1 + 2r(m,k).\cos\theta(m,k). \quad (10)$$

Therefore in order to remove the noise effect precisely, we should consider not only the exact noise autocorrelation $r_{vv}(m,k)$, but also the function $M(r(m,k),\theta(m,k))$ should be calculated for each lag.

2.2.1 Calculation of $M(r(m,k),\theta(m,k))$

The variation of the kernel function $M(r(m,k),\theta(m,k))$ in a frame is drawn in Figure 1. We normalized $|v(m,k)|$ between $0 \sim 1$ and $\theta(m,k)$ changes between $-\pi \sim \pi$ with clean speech amplitude equal to 1.

As it is clear from Figure 1, when the noise amplitude $|v(m,k)|$ is large, changes in $\theta(m,k)$ results in large changes in $M(r(m,k),\theta(m,k))$. By comparing equations (8) and (9), we can see that in (8), $M(r(m,k),\theta(m,k))$ is omitted which means that the effect of noise phase is not considered. Therefore, obviously, the ANS method cannot lead to an

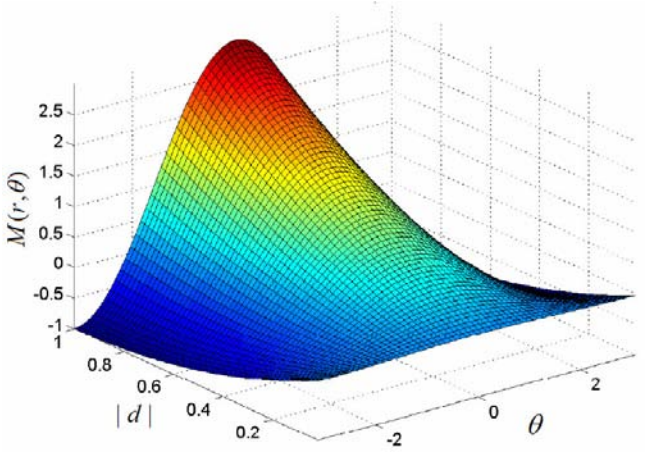

Figure 1 – Variation of $M(r(m,k),\theta(m,k))$ versus $|d(m,k)|$ and $\theta(m,k)$.

appropriate improvement with the changes in SNR. In fact, the problem with ANS method is that it does not take into account the amplitude ratio $r(m,k)$ and phase difference $\theta(m,k)$. From equation (10) we have the noise autocorrelation component as follows

$$z(m,k) = r_{yy}(m,k) - r_{xx}(m,k)$$
$$= |r_{vv}(k)|(\sqrt{r^2(m,k) + 2r(m,k).\cos(m,k) + 1} - r(m,k)) \quad (11)$$

Since we do not know the exact value of phase difference $\theta(m,k)$, the value of

$$\sqrt{r^2(m,k) + 2r(m,k).\cos\theta(m,k) + 1} - r(m,k) \quad (12)$$

cannot be calculated exactly. Hence, we will use its expectation value instead of it, i.e

$$\gamma(r(m,k)) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\{\sqrt{r(m,k)^2 + 2.r(m,k).\cos\theta(m,k) + 1} - r(m,k)\}d\theta \quad (13)$$

This is a function of *r(m,k)* and is shown in Figure 2. Therefore, the noise autocorrelation component is

$$z(m,k) = r_{vv}(k).\gamma(r(m,k)) \quad (14)$$

and autocorrelation of clean speech signal is estimated by

$$r_{xx}(m,k) = r_{yy}(m,k) - z(m,k). \quad (15)$$

For simplicity, according to Figure 2, we replace function $\gamma(r(m,k))$ in one frame of utterance as $\gamma(r)$ and approximate with the following equation which has roughly a similar shape and is found empirically

$$\gamma(r) = exp(a - br), \quad (16)$$

where *a* was set to 1.2 and *b* to 0.45.

Therefore, in our implementations, we have used (16) instead of (13). This new proposed method which has used the function $M(r(m,k),\theta(m,k))$ to consider the cross correlation term between clean speech and noise is named Kernel method.

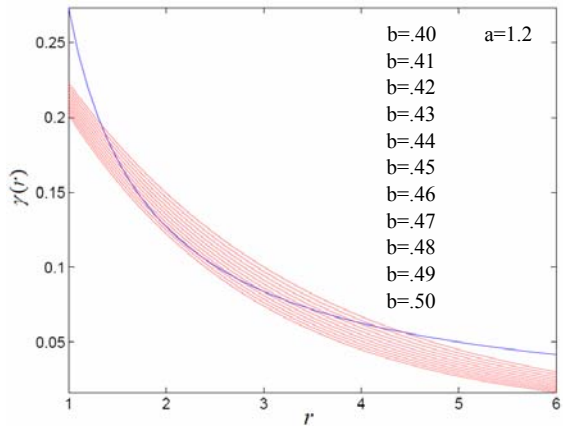

Figure 2 – Function $\gamma(r)$

2.3 Overestimation Parameter in Kernel Method

By subtraction of noise autocorrelation sequence from the autocorrelation of noisy speech, some peaks will be added to the estimated autocorrelation sequence of clean speech signal, which is caused by valleys of the noise autocorrelation sequence. To reduce the effects of these peaks, we use an overestimation parameter.

Therefore, the estimated autocorrelation sequence of noise is modified by multiplying parameter a by $r_{vv}(m,k)$, where a is the overestimation parameter.

2.4 Mean and Variance Normalization of Energy and Cepstrum Vector

As an extra step to further boost the robustness of speech recognition, we have used the signal energy instead of its logarithm in the feature vector and applied normalization to the mean and variances of the cepstral and energy parameters. This approach has led to substantial improvements in the system recognition performance in previous research [10]. Here, we will test its suitability to be combined with our autocorrelation-based approach.

3. PROPOSED ALGORITHM

In this section, we describe the feature extraction algorithm based on our proposed method for the consideration of cross correlation and overestimation parameter plus the mean and variance normalization of energy and cepstal parameters and finally setting of the parameters.

3.1 Inserting Cross Correlation term in ANS Method

As we mentioned, ANS does not consider the cross correlation effect between noise and clean speech signal. Here, we will describe the algorithm of our proposed Kernel method to overcome this problem of ANS. In order to implement the Kernel method we present the following procedure to extract speech features.

1. Frame blocking and pre-emphasizing.
2. Hamming windowing.
3. Calculation of the unbiased autocorrelation sequence of noisy speech signal.

4. Estimation of noise autocorrelation sequence using a few initial frames of each utterance.
5. Calculation of the cross-correlation term between noisy speech signal and noise.
6. Subtracting $M(r(m,k),\theta(m,k))$ times autocorrelation sequence of noise from the autocorrelation sequence of noisy speech signal according to equation (9).
7. Calculation of Fast Fourier Transform (FFT).
8. Calculating the logarithms of mel-frequency filter bin values.
9. Cepstral parameter calculation by applying discrete cosine transform (DCT) to the resulting sequence of step 8.
10. Dynamic cepstral parameter calculations.

Most of the steps in this algorithm are the same as the normal MFCC calculations. Only steps 3 to 6 are newly added steps. These steps include the calculation of the autocorrelation sequence of the noisy signal, estimation of the noise autocorrelation sequence, calculation of the cross correlation between noisy speech and noise, and calculating and inserting $M(r(m,k),\theta(m,k))$ into (9).

3.2 Setting of Parameters

For the estimation of the noise autocorrelation sequence, we used the first 20 frames of the noisy signals from each utterance in Aurora2 corpus. As mentioned in [7], this number of frames has shown the best recognition rate in comparison to other numbers. We used an SNR-adaptive overestimation parameter in our experiments. The SNR was calculated as

$$SNR = 10\log_{10}\frac{\sum_{k=0}^{N-1}|Y(k)|^2}{\sum_{k=0}^{N-1}|\hat{V}(k)|^2}, \qquad (17)$$

where N is the Fourier transform length and $Y(k)$ and $\hat{V}(k)$ are Fourier transforms of noisy signal and noise respectively. After SNR calculation we have used Figure 3 to calculate overestimation parameter a .

4. EXPERIMENTAL EVALUATIONS

Our proposed approach was implemented on Aurora2 recognition task [12]. The pre-emphasis coefficient was set to 0.97. For each speech frame, a 23-channel mel-scale filter-bank was used. The feature vectors for Kernel method were composed of 12 cepstral and log-energy parameters (except for the case shown with MVN where energy was used), together with their first and second order derivatives (39 coefficients in total). All model creation, training and tests in all our experiments have been carried out using HTK [13].

In Figure 4, the results obtained using MFCC, RAS, ANS and our proposed Kernel, Kernel with overestimation parameter named Kernel+OEP and Kernel with overestimation parameter and normalization of mean and Variance of energy and cepstral parameters named Kernel+OEP+MVN are shown.

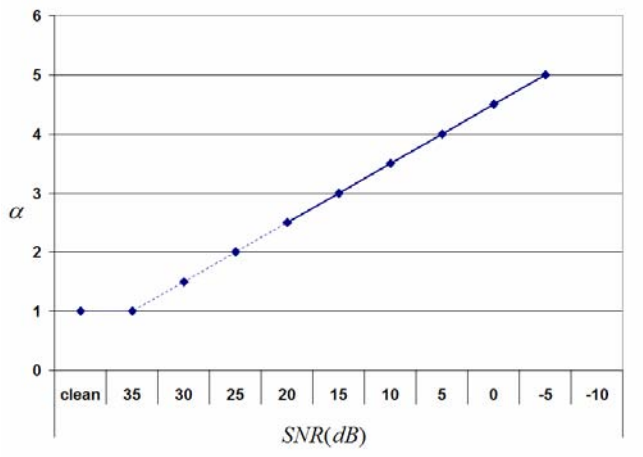

Figure 3 – Variation of a with SNR on Aurora2 corpus.

According to this figure, Kernel has led to better results in comparison to MFCC, RAS and ANS methods for all test sets. Although the amount of improvement is not high, these results indicate the effectiveness of the cross correlation term in the calculation of the estimated autocorrelation of clean speech signal. By applying overestimation parameter on the Kernel method we have obtained better results than Kernel. The best results were obtained by applying mean and variance normalization of energy and cepstral parameters on the Kernel+OEP method. Also Table 1 shows the average recognition rates obtained for each test set of Aurora2 for MFCC, RAS, ANS, Kernel, Kernel+OEP and Kernel+OEP+MVN methods and also percentage of improvements in comparison to the baseline MFCC method.

As shown in Figure 4, the recognition rates using MFCC is seriously degraded in lower SNRs, while RAS, ANS and Kernel are more robust to different noises with Kernel+OEP+MVN outperforming all the others. As seen in Figure 4, in lower SNRs, the recognition rate of our proposed method is much better than MFCC.

5. CONCLUSION

In this paper we presented the results of a modification to a previously proposed successful autocorrelation-based method, where the cross correlation error term and overestimation parameter are included.

As it is clear from the results obtained on Aurora2, the cross correlation term is a rather important parameter in autocorrelation-based feature extraction. Its effectiveness in such approaches has been shown in this paper. Also, overestimation parameter has been found useful in noise autocorrelation estimation. Energy and cepstral mean and variance normalization has also been found useful in combination with the above-mentioned autocorrelation-based approaches. The combination of these approaches has led to substantial improvements in the recognition performance in unmatched conditions, in comparison to the baseline.

Apparently, other methods for cross correlation estimation are also available. Finding the best of such methods can be taken into account in future research.

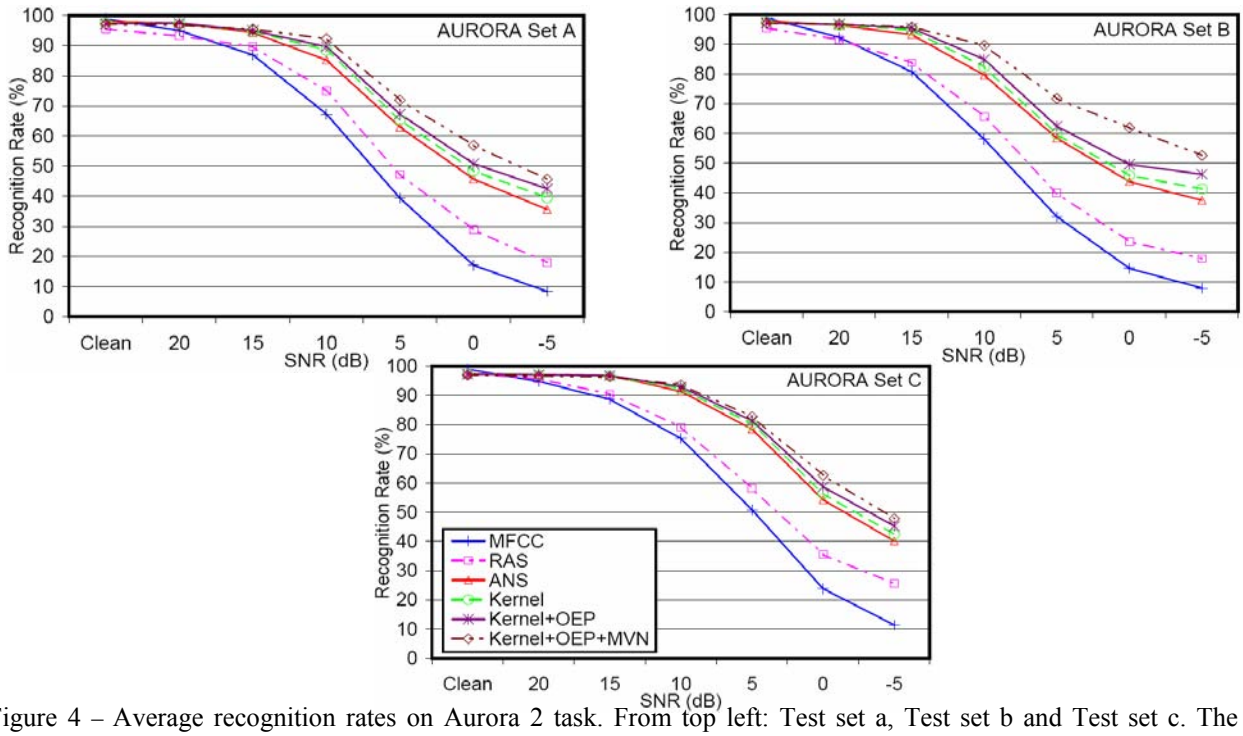

Figure 4 – Average recognition rates on Aurora 2 task. From top left: Test set a, Test set b and Test set c. The results correspond to MFCC, RAS, ANS, Kernel, Kernel+OEP and Kernel+OEP+MVN methods.

Table 1 – Comparison of Average recognition rates for various feature types on three test sets of Aurora 2 task.

Feature type	Average Recognition Rate (%)		
	Set A	Set B	Set C
MFCC	61.13	55.57	66.68
RAS	66.77 (14.51%)	60.94 (12.09%)	71.81 (15.40%)
ANS	77.10 (41.09%)	74.32 (42.20%)	83.61 (50.81%)
Kernel	78.90 (45.72%)	75.88 (45.71%)	84.53 (53.57%)
Kernel+OEP	80.05 (48.68%)	77.86 (50.17%)	85.40 (28.07%)
Kernel+OEP+MVN	82.69 (55.47%)	83.21 (62.21%)	86.42 (59.24%)

REFERENCES

[1] D. Mansour and B.-H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 37, no. 6, pp. 795-804, 1989.

[2] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 5, no.1, pp. 80-84, 1997.

[3] Kuo-Hwei Yuo and Hsiao-Chuan Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Communication*, vol. 28, pp.13-24, 1999.

[4] B.J. Shannon and K.K. Paliwal, "MFCC Computation from Magnitude Spectrum of higher lag autocorrelation coefficients for robust speech recognition," in *Proc. ICSLP 2004,* Jeju.

[5] G. Farahani, S.M. Ahadi and M.M. Homayounpour, "Features based on filtering and spectral peaks in autocorrelation domain for robust speech recognition," *Comput. Speech and Lang* ., Vol. 21, No. 1, pp. 187-205, 2007.

[6] S. Ikbal, H. Misra and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features", in *Proc. ICASSP*, Hong Kong, pp. II-133-136, April 2003.

[7] G. Farahani, S. M. Ahadi, and M. M. Homayounpoor, "Robust Feature Extraction of Speech via Noise Reduction in Autocorrelation Domain," *Lecture Notes in Computer Science 4105*, pp. 466-473, Springer-Verlag, 2006.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing* ASSP 27: 113-120, 1979.

[9] N. W. D. Evans, J. S. D. Mason, W. M. Liu and B. Fauve, "An assessment on the fundamental limitations of spectral subtraction," in *Proc. ICASSP 2006,* Toulouse.

[10] S. M. Ahadi, H. Sheikhzadeh, R. L. Brennan and G. H. Freeman, " An Energy Scheme for Improved Robustness in Speech Recognition," in *Proc. ICSLP*, Jeju, 2004.

[11] K. Onoe, H. Segi, T. Kobayakawa, S. Sato, T. Imai and A. Ando, "Filter Bank Subtraction for Robust Speech Recognition," in *Proc ICSLP*, Colorado, USA, 2002.

[12] H.G. Hirsch, D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000.

[13] The hidden Markov model toolkit available from http://htk.eng.cam.ac.uk.