ROBUST FEATURE EXTRACTION USING AUTOCORRELATION DOMAIN FOR NOISY SPEECH RECOGNITION

Gholamreza Farahani

Department of Electrical Engineering and Information Technology, Iranian Research Organization for Science and Technology (IROST), Tehran, Iran

ABSTRACT

Previous research has found autocorrelation domain as an appropriate domain for signal and noise separation. This paper discusses a simple and effective method for decreasing the effect of noise on the autocorrelation of the clean signal. This could later be used in extracting mel cepstral parameters for speech recognition. Two different methods are proposed to deal with the effect of error introduced by considering speech and noise completely uncorrelated. The basic approach deals with reducing the effect of noise via estimation and subtraction of its effect from the noisy speech signal autocorrelation. In order to improve this method, we consider inserting a speech/noise cross correlation term into the equations used for the estimation of clean speech autocorrelation, using an estimate of it, found through Kernel method. Alternatively, we used an estimate of the cross correlation term using an averaging approach. A further improvement was obtained through introduction of an overestimation parameter in the basic method. We tested our proposed methods on the Aurora 2 task. The Basic method has shown considerable improvement over the standard features and some other robust autocorrelation-based features. The proposed techniques have further increased the robustness of the basic autocorrelation-based method.

KEYWORDS

Autocorrelation, Noise, Robustness, Speech Recognition

1. INTRODUCTION

Mismatch between speech data collected in a laboratory environment, usually used in system training, with those collected in real environments is known to be among the major reasons for speech recognizer performance degradations. Various sources may cause such a mismatch including additive background noise, convolutional channel distortions, acoustic echo and different interfering signals. In this paper, additive background noise is our major concern.

To improve the recognition performance in the presence of additive noise, several approaches have been proposed during the past few decades. While these methods can be very roughly classified into model-based and feature-based, we are more interested in feature-based robust speech recognition.

If one aims to appropriately handle mismatches in the features, he may either try to improve the signal quality before starting to extract recognition features or may try to develop features that are more robust to noise. The first approach is usually known as speech enhancement and is usually dealt with separately from the issue of speech recognition. There are many techniques proposed to solve the speech enhancement problem, most of which concentrate on the spectral domain. On the other hand, several approaches try to extract more noise-robust features for speech recognition. Such methods try to improve recognition performance in comparison to the rather standard features, Mel-Frequency Cepstral Coefficients (MFCC) that have shown good performance in clean-train/clean-test conditions, but deteriorated performance in the cases of mismatch. A very well-known and widely used enhancement method that deals with the signal spectrum is Spectral Subtraction (SS) [1]. Although spectral subtraction is simple in implementation, some levels of success have been observed from its use in combination with speech recognizers. However, this has been limited. Inherent errors in this approach, such as phase, magnitude and cross term errors [2], can lead to performance limitations in enhancement. However, when used in combination with speech recognition systems, some of these errors can be disregarded. Meanwhile some other enhancement methods have been able to achieve more improved performance when used in combination with speech recognizers.

Plenty of research work has been dedicated to extraction of more robust features for speech recognition. One approach, that we are particularly interested in, and has shown some degrees of success in recent works, is the use of autocorrelation in the feature extraction process. Autocorrelation, among its different properties, is known to have a pole preserving property [3]. As an example, if the original signal is modeled by an all-pole sequence, the poles of the autocorrelation sequence will be the same as those of the original signal. Therefore, there exists a possibility of replacing features extracted from the original speech signal with those extracted from its autocorrelation sequence. Consequently, any effort resulting in an improved autocorrelation sequence in the presence of noise could also be helpful in finding more appropriate speech features.

Autocorrelation domain is useful in the different parts related to speech. In Reference [4], different methods of separating voiced and unvoiced segments of a speech signals based on short time energy calculation, short time magnitude calculation, and zero crossing rate calculation on the basis of autocorrelation of different segments of speech signals is introduced. Pitch detection algorithms (PDA) for simple audio signals based on zero-cross rate (ZCR) and autocorrelation function (ACF) in Reference [5] is presented.

Several methods have been reported in autocorrelation domain, leading to more robust sets of features. These methods may be divided into two groups: one dealing with the magnitude of the autocorrelation sequence whilst the other works on the phase of the autocorrelation sequence.

Dealing with the magnitude of the autocorrelation sequence, which is our concern in this paper, among the most successful methods, we can name Differentiated Relative Autocorrelation Sequence Spectrum (DRASS) [6], Short-time Modified Coherence (SMC) [7], One-Sided Autocorrelation LPC (OSALPC) [8], Relative Autocorrelation Sequence (RAS) [9], Autocorrelation Mel Frequency Cepstral Coefficients (AMFCC) [10] and Differentiation of Autocorrelation Sequence (DAS) [11]. Also, it has been shown that the use of spectral peaks obtained from a filtered autocorrelation sequence can lead to a good performance under noisy conditions [12, 13].

In DRASS, autocorrelation will calculate by biased estimator after frame blocking and pre-emphasis. Then after filtering, FFT will calculate and absolute amplitude of differentiated FFT square amplitude will use for Mel scale frequency bank. Finally log of coefficients and cepstrum of them will use as DRASS coefficients.

In SMC, after calculation of autocorrelation with coherence estimation and hamming filtering, the FFT of autocorrelation amplitude is found. Then, applying IFFT, the LPC coefficients are calculated with Levinson-Durbin method and finally the cepstrum of LPC is found as SMC coefficients. In OSALPC, calculation of autocorrelation is carried out by biased estimator and Hamming filtering, the LPC coefficients are calculated using Levinson-Durbin method and the LPC cepstrals found as the final coefficients. Among methods that have made use of the phase of the autocorrelation sequence to obtain a more robust set of features we can name Phase AutoCorrelation (PAC) approach [14] and Autocorrelation Peaks and Phase features (APP) [13]. In this paper, we will consider a few developed autocorrelation-based methods and discuss their approach to achieving robustness. Then we will explain a simple method that can lead to better results in robust speech recognition in comparison to its predecessors in autocorrelation domain. Later, we will discuss the issue of the error terms introduced in this approach due to the estimation of noise autocorrelation sequence. We will show that taking into account the above parameters in the estimation of clean signal autocorrelation sequence can lead to even better system performance.

The remainder of this paper is organized as follows. In Section 2 we will present the theory behind some autocorrelation-based approaches. Section 3 is dedicated to discussion on our Autocorrelation-based Noise Subtraction (ANS) approach and its derivatives. In Section 4, some implementation issues regarding the proposed methods will be addressed. Section 5 includes the experimental results on Aurora 2 task and compares our results with those of the traditional methods such as MFCC and other autocorrelation-based methods. Section 6 will conclude the paper.

2. REVIEW OF THE AUTOCORRELATION-BASED METHODS

In this section, we will describe a few methods in autocorrelation domain. This will give us an appropriate insight to the advantages and disadvantages of using autocorrelation in robust feature extraction.

2.1. Formulation of Clean and Noisy Speech Signals and Noise in Autocorrelation Domain

We start by explaining the relationship between the autocorrelation sequences of clean and noisy signals and noise. Assuming *v(m,n)* to be the additive noise and *x(m,n)* clean speech signal , the noisy speech signal, *y(m,n),* could be written as

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

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. Note that in this paper, as our goal is suppression of the effect of additive noise fromnoisy speech signal, the channel effects not considered. If *x(m,n)* and *v(m,n)* are considered uncorrelated, then the autocorrelation of the noisy speech signal can be written as

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

where $r_{yy}(m,k)$, $r_{xx}(m,k)$ and $r_w(m,k)$ are the short-time one-sided autocorrelation sequences of the noisy speech, clean speech and noise respectively and k is the autocorrelation sequence index within each frame. The one-sided autocorrelation sequence of noisy speech signal may be calculated using an unbiased estimator, i.e.

$$r_{yy}(m,k) = \frac{1}{N-k}\sum_{i=0}^{N-1-k} y(m,i)y(m,i+k) \tag{3}$$

Meanwhile, although reasonable in practice, considering the clean speech signal, $x(m,n)$, and noise, $v(m,n)$, completely uncorrelated may not always be an accurate assumption. We will discuss this issue later. In a more general case, equation (2) should be written as

$$\begin{aligned} r_{yy}(m,k) &= r_{xx}(m,k) + r_{vv}(m,k) + E\{x(m,k).v^*(m,k)\} + E\{x^*(m,k).v(m,k)\} \\ &= r_{xx}(m,k) + r_{vv}(m,k) + r_{xv}(m,k) + r_{vx}(m,k) \qquad 0 \le m \le M-1 \qquad 0 \le k \le N-1 \end{aligned} \tag{4}$$

where $r_{xv}(m,k) = E\{x(m,k).v^*(m,k)\}$ and $r_{vx}(m,k) = E\{v(m,k).x^*(m,k)\}$ are the cross correlation terms between the clean speech signal and noise.

If the noise autocorrelation sequence is assumed relatively constant across frames, we can find an estimate of $r_w(m,k)$ using the non-speech sections of an utterance, specified for example, by a voice activity detector (VAD) or the initial normally non-speech periods and denote it as $\hat{r}_{vv}(k)$. Then we will have

$$r_{yy}(m,k) = r_{xx}(m,k) + \hat{r}_{vv}(k) + r_{xv}(m,k) + r_{vx}(m,k) \qquad 0 \le m \le M-1 \quad 0 \le k \le N-1. \tag{5}$$

Obviously, an assumption of $v(m,n)$ having zero mean and being uncorrelated with $x(m,n)$ will reduce the terms $r_{xv}(m,k)$ and $r_{vx}(m,k)$ to zero [15]

2.2. Autocorrelation-based Methods for Robust Feature Extraction

Recently, several autocorrelation-based methods have been proposed, where usually, the speech signal and noise were considered uncorrelated. We will describe some of these methods here, in order to get some insight on how autocorrelation properties may be used to achieve robustness.

2.2.1. Relative Autocorrelation Sequence (RAS)

As explained in reference [9], this method assumed the noise as stationary and uncorrelated to the speech signal. Therefore, the relationship between the autocorrelations of noisy and clean signals and noise could be written as

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{vv}(k), \qquad 0 \le m \le M-1 \quad 0 \le k \le N-1. \tag{6}$$

If the noise part could be considered stationary, differentiating both sides of equation (6) with reference to the frame index m would remove the effect of noise from the results, i.e.

$$\frac{\partial r_{yy}(m,k)}{\partial m} = \frac{\partial r_{xx}(m,k)}{\partial m} + \frac{\partial r_{vv}(k)}{\partial m} \cong \frac{\partial r_{xx}(m,k)}{\partial m} = \frac{\sum_{t=-L}^{L} t.r_{yy}(m+t,k)}{\sum_{t=-L}^{L} t^2}, \quad 0 \le m \le M-1 \quad 0 \le k \le N-1 . \quad (7)$$

The right side of equation (7) is equal to filtering on the one-sided autocorrelation sequence by a high-pass FIR filter, where L is the length of the filter. This high pass filter (differentiation), named RAS filter, was used to suppress the effect of noise in the autocorrelation sequence of the noisy signal. Therefore, this method is appropriate for noises which have slow variations in the autocorrelation domain, i.e. could be considered as relatively stationary.

After calculating one-sided autocorrelation sequence and differentiating both sides of equation (6) with respect to *m*, the autocorrelation of noise was removed, i.e. differentiation of noisy speech signal is equal to the differentiation of clean speech signal with respect to the frame index, *m*, in autocorrelation domain. Obviously, this filtering will also have some slight negative effects on the lower modulation frequencies of speech. However, this has been found to be quite small (refer to Section 5 for RAS performance in clean speech conditions).

2.2.2. Autocorrelation Mel-frequency Cepstral Coefficients (AMFCC)

In this approach [10], the MFCC coefficients were extracted from the noisy signal autocorrelation sequence after removing some of its lower lag coefficients. These lower lag coefficients were shown to have the highest influence on the noisy signal for many noise types, including those with least correlations among frames. The lag threshold value used was 3 msec. and was set by finding the first valley in the absolute autocorrelation function found over TIMIT speech frames.

As reported in reference [10], this method works well for car and subway noises in Aurora 2 task, but not for babble and exhibition noises. The reason was believed to be wider autocorrelation functions of the latter ones. However, for some other noise types, such as babble, they are spread out in different lags. Therefore, the main reason for limited success of AMFCC in noises such as babble and exhibition is that the noise autocorrelation properties are more similar to those of the speech signal, which makes their separation difficult.

2.2.3. Differentiation of Autocorrelation Sequence (DAS)

This algorithm combines the use of the enhanced autocorrelation sequence of the noisy speech and the spectral peaks found from the autocorrelation sequence, as they are known to convey the most important information of the speech signal [11].

In this method, in order to preserve speech spectral peaks, spectral differentiation has been used. With this differentiation, the flat parts of the spectrum were almost removed and each spectral peak was split into two, one positive and one negative. The differential power spectrum of the noisy signal was defined as

$$Diff_Y(k) \approx \sum_{l=-Q}^{P} a_l . Y(k+l), \qquad 0 \le k \le K-1 \qquad (8)$$

where P and Q are the orders of the difference equation, a_l are real-valued coefficients and K is the length of FFT (on the positive frequency side) [16]. The differentiation mentioned in equation (8) can be carried out in several ways, as discussed in reference [16]. The simple difference had shown the best results and therefore was used in reference [11], i.e.

$$Diff_Y(k) = Y(k) - Y(k+1) \tag{9}$$

The procedure of feature extraction was carried out after high-pass filtering (as in equation (7)) and peak extraction (as in equation (9)). As explained earlier for RAS, this filtering can suppress the effect of slowly varying noises and also attenuate the effect of slow variation noise on the speech signal. The spectral peaks were then extracted through differentiation of the spectrum found using the filtered autocorrelation sequence, leading to better suppression of the noise effect. Finally, an MFCC-like feature set was extracted and used in recognition experiments.

2.2.4. Spectral Peaks of Filtered Higher-lag Autocorrelation Sequence (SPFH)

This method was proposed to overcome the main drawback of AMFCC, i.e. its inability to deal with noises that have autocorrelation components spread out over different lags [17].

In SPFH, after frame blocking and pre-emphasis of the noisy signal, the autocorrelation sequence of the frame signal was obtained as in equation (3) and its lower lags were removed. A FIR high-pass filter, similar to RAS filter, was then applied to the signal autocorrelation sequence to further suppress the effect of noise, as in equation (7). Then, Hamming windowing and short-time Fourier transform were carried out and the differential power spectrum of the filtered signal was found using equation (9). Since the noise spectrum may, in many occasions, be considered flat, in comparison to the speech spectrum, the differentiation either reduces or omits these relatively flat parts of the spectrum, leading to even further suppression of the effect of noise. The final stages included applying the resultant magnitude of the differentiated autocorrelation-derived power spectrum to a conventional mel-frequency filter-bank and passing the logarithm of the outputs to a DCT block to extract a set of cepstral coefficients per frame.

In fact, the SPFH method tried to attenuate the effect of noise after preserving higher lags of noisy autocorrelation sequence by high-pass filtering, as in equation (7), and preserving spectral peaks, as in equation (9), i.e. similar to DAS.

3. NOISE SUBTRACTION IN AUTOCORRELATION DOMAIN

3.1. Autocorrelation-based Noise Subtraction (ANS)

As an ideal assumption, we can consider the autocorrelation of noise as a unit sample at the origin and zero at other lags. Therefore that portion of noisy speech autocorrelation sequence which is far enough from the origin will have the same autocorrelation as clean speech signal. This ideal assumption is of course only true for white noise and for real environmental noises, components in lags other than zero are also available.

Investigations showed that there exist some major autocorrelation components for these noises concentrated around the origin. This was the reason for introducing AMFCC and SPFH methods mentioned earlier. However, as these methods drop the lower lags of the autocorrelation sequence

of the noisy speech signal to suppress the effect of noise, they are not useful for the cases where important components are seen in higher autocorrelation lags of the noise, i.e. above 20 to 25 samples. In such cases, AMFCC approach not only does not completely suppress the effect of noise, but also removes some probably useful lower lag portions of the autocorrelation sequence of the speech signal. As an alternative to such methods, we follow a newer approach. Here, in place of removing the lower lag autocorrelation components of the noisy signal, we try to estimate the noise autocorrelation sequence and deduct it from the noisy signal autocorrelation sequence. This is conceptually similar to the well-known spectral subtraction with the exception that it is not magnitude spectrum, but to the autocorrelation sequence [17]. An instant advantage is that there is no need to deal with phase issue.

In reference [17], the average autocorrelation of a number of non-speech frames of the utterance is used as an estimate of the noise autocorrelation sequence. We write this as

$$\hat{r}_{vv}(k) = \frac{\sum_{i=0}^{P-1} r_{yy}(i,k)}{P}, \qquad 0 \le k \le N-1 \tag{10}$$

where P is the number of non-speech frames of the utterance used and $\hat{r}_{vv}(m,k)$ is the noise autocorrelation estimate.

Therefore, we may write the estimate of the autocorrelation sequence of the clean speech signal as

$$\hat{r}_{xx}(m,k) = r_{yy}(m,k) - \hat{r}_{vv}(m,k). \tag{11}$$

In order to estimate the noise autocorrelation in ANS method, a voice activity detector (VAD), or the initial silence of the speech utterances can be used. Note that procedures similar to many other widely-used noise estimation methods could also be used here.

3.2. The Cross Correlation Term

Figure 1 displays the autocorrelation sequences for two examples of clean speech, noise and noisy speech signals with the noises being babble and factory, extracted from the NATO RSG-10 corpus [18], as well as the sum of autocorrelation sequences of speech and noise. One should expect the speech signal and noise, in most circumstances, to be completely uncorrelated However, in this case, according to figure 1, the autocorrelation sequence of the noisy speech is not equal to the sum of those of clean speech and noise. In order to be able to have a more accurate estimate of the speech signal autocorrelation, one needs to consider some correlation among speech and noise signals to compensate for this difference. It should be noted that this difference is in fact due to the short-time nature of our analysis, as the simple form of additive autocorrelation mentioned in equation (2) is only possible when an infinitely long signal is considered in the analysis [19]. We have used the two following approaches in order to consider the cross correlation term in autocorrelation calculations:

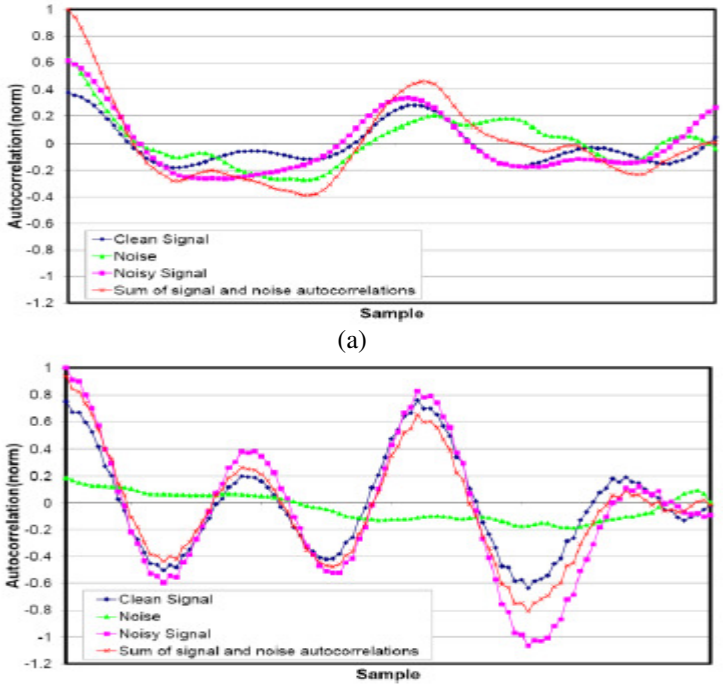


Figure 1. Sample autocorrelation sequences of the clean speech, noisy speech and noise as well as sum of the autocorrelations of clean speech signal and noise with (a) babble noise and (b) factory noise, with an SNR of 10dB.

3.2.1. 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. If we also consider the above mentioned correlation between the clean speech signal and noise, the relationship between the autocorrelations mentioned in equation (6) should change to:

$$r_{yy}^2(m,k) = r_{xx}^2(m,k) + r_{vv}^2(k) + 2.r_{xx}(m,k).r_{vv}(k).\cos\theta(m,k) \quad (12)$$

where $\theta(m,k)$ is the instantaneous phase difference between clean speech signal autocorrelation, $r_{xx}(m,k)$, and noise autocorrelation, $r_{vv}(k)$. From equation (12), we will have [20]

$$r^2{}_{xx}(m,k) = r^2{}_{yy}(m,k) - r^2{}_{vv}(k).(1 + 2r(m,k).\cos\theta(m,k)) = r^2{}_{yy}(m,k) - M(r(m,k),\theta(m,k)).r^2{}_{vv}(k) \quad (13)$$

where

$$r(m,k) = \frac{|r_{xx}(m,k)|}{|r_{vv}(k)|} \quad (14)$$

$$M(r(m,k),\theta(m,k)) = 1 + 2r(m,k).\cos\theta(m,k)$$

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

The variation of the kernel function $M(r(m,k),\theta(m,k))$ in a frame is drawn in figure 2. We normalized $|r_w(k)|$ between $0\sim1$ and named it $|d(k)|$. Also $\theta(m,k)$ changes between $-\pi$ to π with clean speech amplitude equal to 1.

As it is clear from figure 2, when the noise autocorrelation amplitude $|r_w(k)|$ is large, changes in $\theta(m,k)$ result in large changes in $M(r(m,k),\theta(m,k))$.

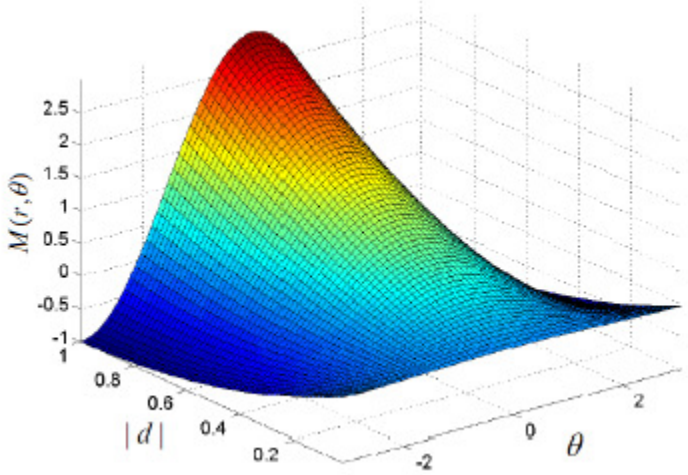


Figure 2. Variation of $M(r(m,k),\theta(m,k))$ versus $|d(k)|$ and $\theta(m,k)$.

In the following equation we have the noise autocorrelation component as [20]

$$z(m,k)=r_{yy}(m,k)-r_{xx}(m,k)=|r_w(k)|(\sqrt{r^2(m,k)+2r(m,k).\cos\theta(m,k)+1}-r(m,k)) \quad (15)$$

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

$$\sqrt{r^2(m,k)+2r(m,k).\cos\theta(m,k)+1}-r(m,k) \quad (16)$$

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

$$\gamma(r(m,k))=\frac{1}{2\pi}\int_{-\pi}^{\pi}\{\sqrt{r(m,k)^2+2.r(m,k).\cos\theta(m,k)+1}-r(m,k)\}d\theta \quad (17)$$

This is a function of *r(m,k)* and is shown in figure 3. Therefore, the noise autocorrelation component is

$$z(m,k)=r_{ww}(k).\gamma(r(m,k)) \quad (18)$$

and the autocorrelation of clean speech signal is estimated by

$$r_{xx}(m,k)=r_{yy}(m,k)-z(m,k). \quad (19)$$

For the sake of simplicity, according to figure 3, we change the function $\gamma(r(m,k))$ in one frame of utterance to $\gamma(r)$ and replace it with an approximate value found using the following equation, which has roughly a similar shape and is found empirically

$$\gamma(r) = \exp(a - br), \tag{20}$$

where a was set to 1.2 and b to 0.45 in our experiments.
Therefore, in our implementations, we have used equation (19) instead of equation (11). We named this method, which has used the function $M(r(m,k),\theta(m,k))$ to consider the cross correlation term between clean speech and noise, as Kernel method.

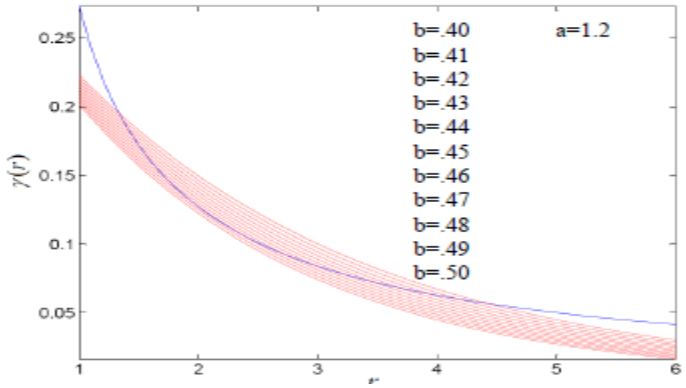


Figure 3. Function $\gamma(r)$. The bunch of curves represents the estimation of the function according to different values of b. The other curve represents the real values of $\gamma(r)$.

3.2.2 Autocorrelation Averaging

We used autocorrelation averaging as an alternative way for reducing the observed correlation effect between noise and clean speech signal. We remind the reader that, as already mentioned, this correlation might even solely be the result of autocorrelation analysis on finite-duration signals. In reference [21], it was shown that a smoothing approach can help in spectral subtraction to overcome the speech/noise correlation problems. The reason is that the probability density function (pdf) of the cosine of the angle between speech and noise vectors has been shown to have a minimum at zero, while smoothing leads to a pdf with a maximum at zero and smaller variances with larger numbers of frames taking part in smoothing 1. As a result, as will be further explained later, ignoring the term including $\cos(\theta)$,i.e assuming $\cos(\theta) = 0$, would be less harmful after smoothing. We define the average of the noisy autocorrelation sequence as

$$\overline{r_{yy}}(m,k) = \sum_{i=0}^{T-1} b_i r_{yy}(m-i,k), \quad \sum_{i=0}^{T-1} b_i = 1 \tag{21}$$

i.e. weighted averaging of the noisy speech autocorrelation on T frames where b_i is a weighting parameter larger than 0 and less than or equal to 1.

---

[1]A more detailed discussion on this issue can be found in [20].

By replacing $r_{yy}(m-i,k)$ in equation (21) with the value found in equation (4) we have :

$$\overline{r_{yy}(m,k)} = \sum_{i=0}^{T-1} b_i . r_{yy}(m-i,k) = \sum_{i=0}^{T-1} b_i . r_{xx}(m-i,k) + \sum_{i=0}^{T-1} b_i . r_{vv}(m-i,k) + E\{\sum_{i=0}^{T-1} b_i . x(m-i,k).v^*(m-i,k)\} + E\{\sum_{i=0}^{T-1} b_i . x^*(m-i,k).v(m-i,k)\} \quad (22)$$

If the variations in noise and speech could be assumed negligible during a period T, we can write

$$\sum_{i=0}^{T-1} b_i . r_{xx}(m-i,k) \approx r_{xx}(m,k) , \quad (23)$$

$$\sum_{i=0}^{T-1} b_i . r_{vv}(m-i,k) \approx r_{vv}(m,k) , \quad (24)$$

$$E\{\sum_{i=0}^{T-1} b_i . x(m-i,k).v^*(m-i,k)\} \approx E\{|x(m,k)||v(m,k)|.\sum_{i=0}^{T-1} b_i . \cos\theta(m-i,k)\} , \quad (25)$$

and

$$E\{\sum_{i=0}^{T-1} b_i . x^*(m-i,k).v(m-i,k)\} \approx E\{|x(m,k)||v(m,k)|.\sum_{i=0}^{T-1} b_i . \cos\theta(m-i,k)\} . \quad (26)$$

Setting the value of the parameters T and b_i will be discussed in the parameter settings section (Section 4.3) and we will see that with the values used for T, the above mentioned assumption holds.

It was shown in reference [21] that if the phase differences between the speech and noise in successive frames are assumed to be uncorrelated, the pdfs of the summation terms in equations (25) and (26), depending on the value of T, would peak at zero and have a standard deviation of $1/\sqrt{2T}$. Therefore, the above two terms may be considered as almost zero and equation (22) would be rewritten as

$$\overline{r_{yy}}(m,k) \approx r_{xx}(m,k) + r_{vv}(m,k) \quad (27)$$

By replacing $r_{yy}(m,k)$ with $\overline{r_{yy}}(m,k)$ in equation (11) we have

$$\hat{r}_{xx}(m,k) = \overline{r_{yy}(m,k)} - \hat{r}_w(m,k) \approx r_{xx}(m,k) + r_w(m,k) - \hat{r}_w(m,k) \quad (28)$$

Therefore, if we estimate the autocorrelation sequence of noise, $\hat{r}_{vv}(m,k)$, more accurately, our estimate of the clean speech signal would also be more accurate. The above mentioned process will also have a slight effect on the speech signal. However, as the results of the application of this method on the clean speech show (Section 5), this effect is negligible.

We will call this approach autocorrelation-based noise subtraction with smoothing (ANSS). Details of setting of the length of averaging window in this approach will be discussed in the parameter setting Section 4.3.

3.3 ANS versus Spectral Subtraction

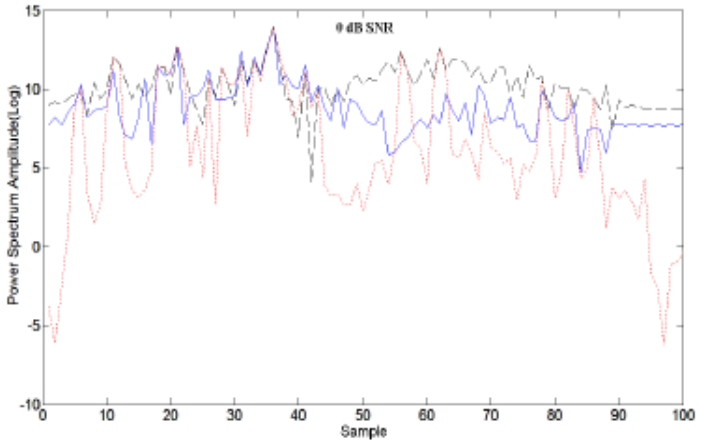
Due to the similarity of ANS and spectral subtraction (SS) in concept, in this section, we would like to make a comparison between the two methods. The first, and by far the most important, difference between these two methods is that the subtraction in SS takes place in spectral domain whereas for ANS, the subtraction is carried out in the autocorrelation domain (time domain). Note that in the implementation of spectral subtraction, reported in this section, the overestimation factor is set equal to that used for ANS and the flooring parameter was set to 0.002.

Although traditional spectral subtraction suffers from a few problems that affect the quality of enhanced speech, the important source of distortion in this method is known to be the negative values encountered during subtraction that should be mapped to a spectral floor [22]. This nonlinear mapping causes an effect that is usually known as musical noise and is always associated with the basic spectral subtraction method.

In ANS, as the subtraction is carried out in autocorrelation domain, negative and positive values are not treated differently, and therefore, there is no need for flooring or other non-linear mappings. In fact, problems associated with non-linearity are not encountered anymore and inaccuracies in speech spectral estimates are only due to errors in noise autocorrelation estimation and its associated problems.

Figure 4 displays the power spectra of a frame of an utterance of the word "one", uttered by a female speaker and contaminated with train station noise at 0dB and 10dB SNRs. This utterance is extracted from test set A of the Aurora 2 task. In this figure, the power spectra of signal after the application of ANS and spectral subtraction are shown. As it is clear, the power spectrum extracted after the application of ANS to the noisy speech closely follows the peaks and valleys of the clean spectrum while the SS-treated one has a more different appearance.

The normalized average spectral errors of both methods have also been shown in Table 1. Apparently, the Root Mean Square Error (RMSE) of ANS is much less than that of spectral subtraction.



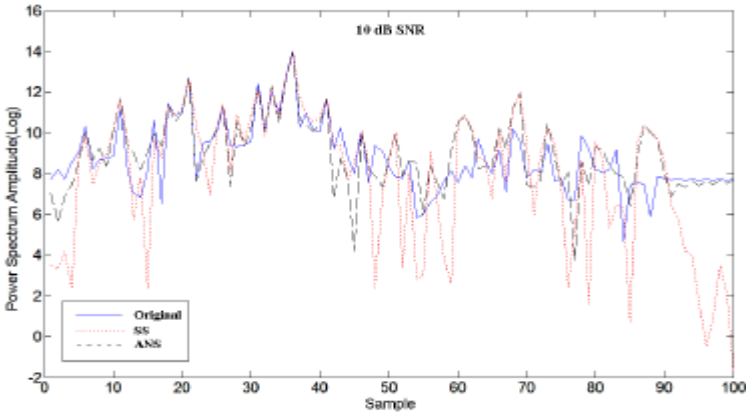


Figure 4. Log power spectrum of a speech frame of 'FAK_1B.08' utterance from test set A of Aurora 2 task contaminated with subway noise at 0dB and 10dB SNRs in logarithmic scale.

Table 1. Normalized Average of Spectral Subtraction and ANS Spectrum Errors (RMSE Criteria) on Test Set A of Aurora 2 Task

Method	Average of Spectrum error					
	20dB	15dB	10dB	5dB	0dB	-5dB
SS	332	378	488	653	982	1500
ANS	22.4	28.1	39.3	55.9	87.9	142

4. IMPLEMENTATION ISSUES IN PROPOSED ALGORITHMS

In this section we will discuss a number of implementation issues regarding our proposed methods. Also in this section, we will consider the overestimation parameter to enable us better estimate the noise autocorrelation sequence.

4.1. Considering Cross Correlation

To consider the cross correlation term, we have implemented two different methods, as discussed in Subsections 3.2.1 and 3.2.2. The procedure for feature extraction in our proposed methods is as follows.

a) Frame Blocking and pre-emphasis.
b) Hamming windowing.
c) Calculation of unbiased autocorrelation sequence of noisy speech signal.
d) Estimation of noise autocorrelation sequence in each utterance and subtracting it from the speech signal autocorrelation sequence in each frame of the utterance (More details of parameter settings will be found in Section 3).
e) Kernel function computation (see Subsection 3.2.1) or autocorrelation averaging (see Subsection 3.2.2).
f) Inserting cross correlation term in the estimation of the autocorrelation of the clean speech.
g) Fast Fourier Transform (FFT) calculation.
h) Calculation of the logarithms of Mel-frequency filter bank outputs.

i) Application of DCT to the sequence resulting from pervious step.
j) Calculation of the feature vectors including 12 cepstral and a log-energy parameter and their first and second order dynamic parameters.

In these algorithms, almost all the steps are rather straightforward. Only steps e) and f) are added to our implementation of ANS, which are related to inclusion of the cross correlation term. The accuracy of the cross correlation term estimation would be crucial at this stage. The results of our implementations will be given in Section 5.

4.2. Considering Overestimation Parameter

Since our algorithm is applied to the autocorrelation of the noisy signal, the flooring parameter used in spectral subtraction will not be needed in the application of our algorithm. The reason is that flooring in spectral subtraction is usually needed to remove the negative spectral values, while this would not be a problem in autocorrelation domain. As shown in Figure 5, in the autocorrelation sequence of noise, valleys and peaks may be observed whose lag locations and magnitudes might vary from one frame to another.

Although smoothed to some extent, such perhaps unrealistic peaks and valleys might still show up in our estimate of the noise autocorrelation sequence. By subtracting the noise autocorrelation sequence from that of the noisy speech, some peaks and valleys will be added to the estimated clean speech autocorrelation sequence, resulted from valleys and peaks in the estimated noise autocorrelation sequence. In order to decrease the effects of these peaks and valleys, we have used an overestimation parameter by modifying the ANS equation to

$$r_{yy}(m,k) = r_{xx}(m,k) + \alpha \hat{r}_w(k), \tag{29}$$

where $\alpha \geq 1$ is the overestimation parameter. Note that when $\alpha = 1$, equation (29) is identical to the equation used for ANS. Apparently, having $\alpha > 1$ leads to some attenuation in the peaks of the estimated clean speech signal autocorrelation, due to increase in the last term of equation (29). Various values of α were tested to get the best result on the Aurora 2 task.

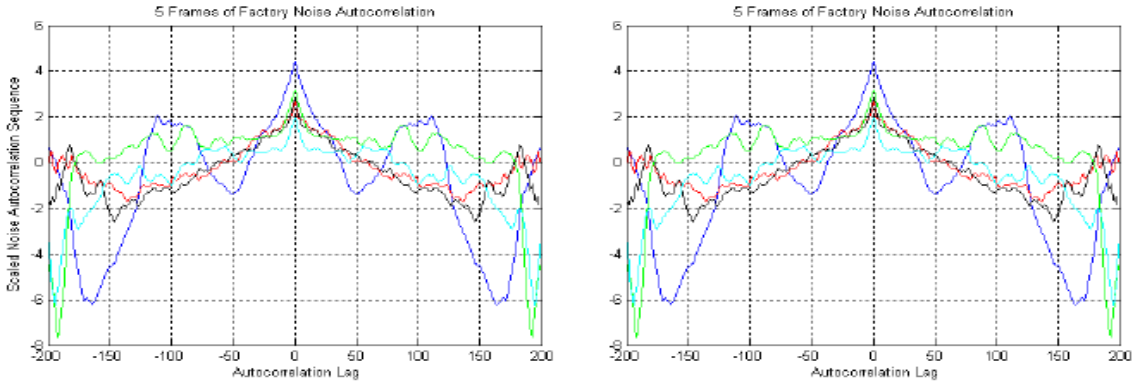


Figure 5. Autocorrelation sequences for 5 consecutive frames of factory noise.

In order to reduce the speech distortions caused by large values of α, we have changed this parameter with SNR [23]. The SNR was calculated frame by frame as explained in the parameter

setting Section 4.3. Figure 6 shows the trend of change we used for parameter α with SNR. Clearly, with increasing SNR, the values of α should decrease and vice versa. The trend of this change was set to linear, as shown in the figure, according to changes observed in system recognition performance in practice [17]. We tested the proposed method with/without taking into account the signal/noise cross correlation. If we consider the issue of cross correlation, as explained in Section 3.2.1, together with the overestimation parameter, the following relationship for clean speech signal estimation will result

$$\hat{r}_{xx}(m,k) = r_{yy}(m,k) - \alpha.\hat{r}_{vv}(k) - \gamma.\hat{r}_{yv}(m,k)\,. \tag{30}$$

Meanwhile, considering the cross correlation term as in Section 3.2.2, together with the overestimation parameter, we will have the following equation, which gives an approximate value of the speech signal.

$$\hat{r}_{xx}(m,k) = \overline{r_{yy}}(m,k) - \alpha.\hat{r}_{vv}(k) \approx r_{xx}(m,k) + r_{vv}(m,k) - \alpha.\hat{r}_{vv}(k) \tag{31}$$

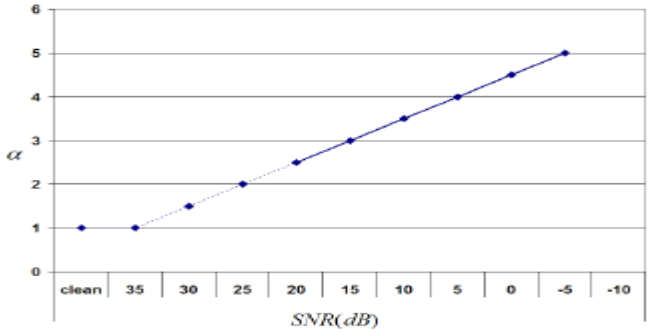


Figure 6. Change in the parameter α with SNR on Aurora 2 task.

4.3. Parameter Settings

In our implementation of RAS, the length of the filter was set to L=2 according to reference [9]. Also the duration for lower lag elimination in the AMFCC method was set to 2.5 ms (20 samples in 8 kHz sampling rate for Aurora 2 task) similar to reference [10]. The same duration was also used for SPFH implementation [16]. In order to estimate the noise autocorrelation sequence, in all our experiments, we have used 20 initial frames of each utterance, considering them as non-speech sections. As shown in reference [24], this number of frames resulted in best recognition rates on Aurora 2 task.

In the implementation of ANSS, in order to get the best results, we have tried different numbers of frames (T in equations (21) to (26)) for averaging. Figure 7 shows the results. As depicted, the grand average recognition rates on the three sets of Aurora 2 task have shown the best performance with 3 frames used in autocorrelation averaging. Therefore, in our experiments, we have used this number for noisy speech autocorrelation averaging. Regarding b_i, in our experiments, simple averaging was carried out. In the implementations using overestimation parameter, this parameter was changed as a function of SNR in each frame. An estimate of SNR in each frame was found as

$$SNR = 10\log_{10}\frac{\sum_{k=0}^{N-1}|Y(k)|^2}{\sum_{k=0}^{N-1}|\hat{V}(k)|^2} \tag{32}$$

where N is the FFT length, $Y(k)$ is the spectrum of the noisy speech signal and $\hat{V}(k)$ is the FFT of the first few frames of the noise autocorrelation sequence estimation. After calculating SNR, we found the overestimation parameter as shown in figure 6.

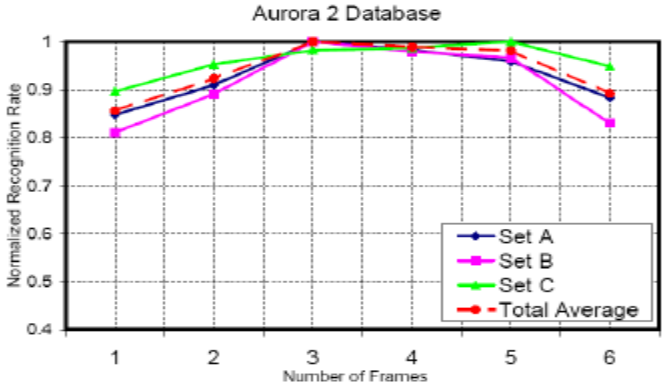


Figure 7. Normalized recognition rates for test sets of Aurora 2 task versus the number of frames used in noise autocorrelation sequence averaging and their grand average.

5. EXPERIMENTS

In this section, we will describe the data used and procedures followed in our experiments. Our implementations include some of the previous methods for comparison purposes as well as our proposed approaches.

5.1. Data

The experiments were carried out on Aurora 2 task [25]. The features in this case were computed using 25 msec. frames with 10 msec. of frame shifts. 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 proposed methods were composed of 12 cepstral and a log-energy parameter, together with their first and second order derivatives. All model creation, training and tests in all our experiments have been carried out using the standard Hidden Markov model toolkit [26] with 16 states and 3 mixture components per state. The HMMs were trained in clean condition, i.e. with clean training data.

5.2. Implementation Results using Cross Correlation Terms

The setting of our parameters is as described in 4.3. Figure 8 includes ANS, Kernel and ANSS recognition results on the Aurora 2 data. Also, for comparison purposes, the results of baseline MFCC, together with RAS, AMFCC and MFCC-SS are included. RAS and AMFCC were chosen as two of the most successful autocorrelation-based methods. Also note that the parameters used in MFCC-SS are the same used in the implementation of spectral subtraction explained in Section 2.3. While the results of ANS, Kernel and ANSS show considerable improvement over the

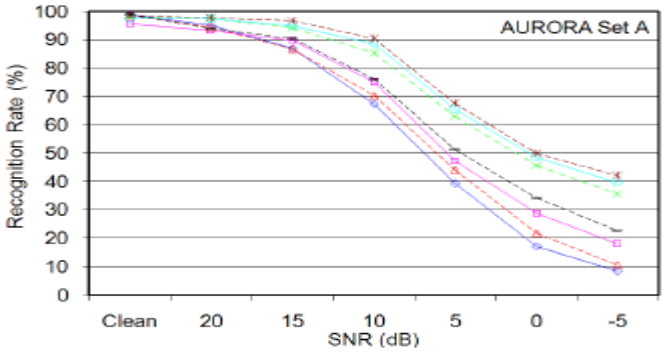
baseline MFCC in noisy conditions, ANSS has shown superior performance in comparison to ANS and Kernel methods. In fact, ANSS has performed quite well, outperforming the standard MFCC with a very large margin, especially in lower SNRs, reaching a value of up to 35% absolute reduction in word error rate. In comparison to ANS, which itself performs satisfactorily in noisy conditions, the higher performance of ANSS is noticeable. A prompt conclusion could be that including the effect of noise-signal cross correlation in autocorrelation-based noise subtraction method can further improve the performance boundaries of this method. This is indicative of the effectiveness of inserting the cross correlation parameter into the autocorrelation calculation of noisy speech signal.

5.3. Implementation Results of Applying Overestimation Parameter

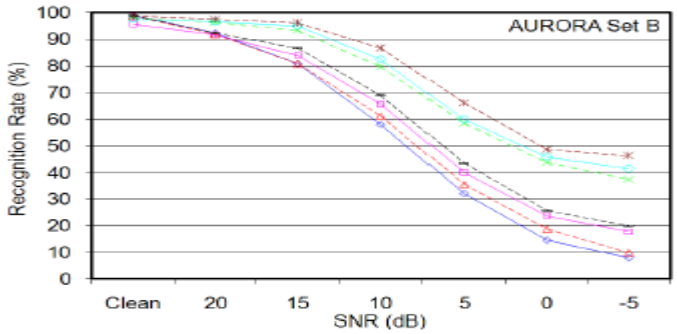
The results of including the overestimation parameter α into clean speech autocorrelation estimation procedure will be reported here. Figure 9 depicts our recognition results on Aurora 2 Task. The naming conventions for our methods are as before with OEP being added to indicate the inclusion of the overestimation parameter in the implementation. As it is clear, the application of overestimation has led to improvements in the system recognition performance in almost all cases. This indicates the potential of the overestimation parameter in improving autocorrelation-based noise subtraction.

5.4. Comparison of the Discussed Methods

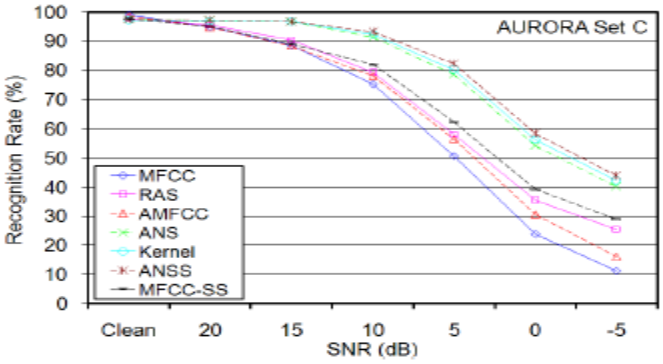
In order to reach to an overall conclusion on different methods discussed, we wish to compare the performances of all the mentioned methods on the specified task. Furthermore, as mentioned in reference [27], using the normalized energy instead of the logarithm energy, together with mean and variance normalization of the cepstral parameters, could lead to improvement in the speech recognition performance in noisy conditions. Therefore, we have also applied this technique which has further improved the recognition rate of our best method discussed, ANSS+OEP. Table 2 shows the average recognition rates of all these methods on the Aurora 2 task. As usual in Aurora 2 result calculations, the -5dB and clean results are not included in the averaging. Furthermore, the percentage of relative improvement of each method in comparison to the baseline MFCC is also mentioned. We have also included two other test results in this table; MFCC enhanced with spectral subtraction (MFCC-SS) and mean subtraction, variance normalization and ARMA filtering (MVA) [28]. The former is meant to show the performance improvement obtained by spectral subtraction, as a basic enhancement approach on this task while the latter is just added as a rather simple method known to perform among the best in robust speech recognition. The implementation procedure was exactly similar to our other tests. Also, in this table, for comparison purposes, the results obtained from the application of ETSI Extended Advanced Front-End [29] on the Aurora2 corpus are reported. This is a standard front-end which uses sophisticated enhancement approaches to improve the quality of the extracted features.



(a)

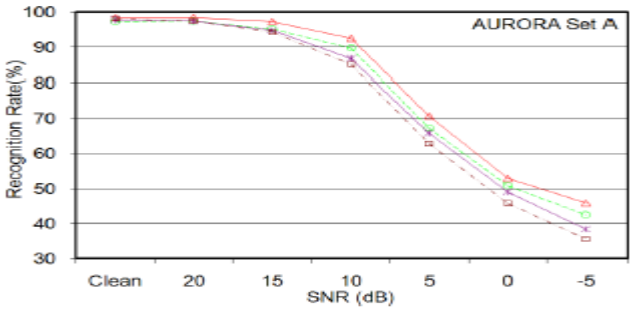


(b)

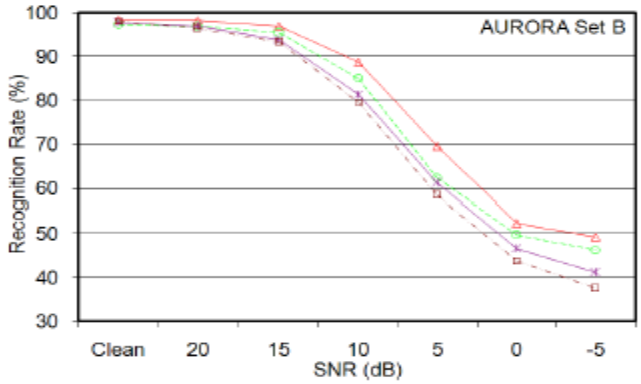


(c)

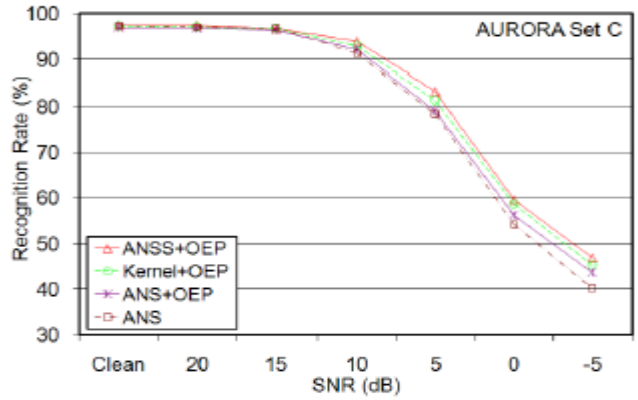
Figure 8. Average recognition rates of MFCC, RAS, AMFCC, ANS, Kernel, ANSS and MFCC-SS on Aurora 2 task. (a) Test set A, (b) Test set B and (c) Test set C.



(a)



(b)



(c)

Figure 9. Average recognition rates of ANSS+OEP, Kernel+OEP, ANS+OEP and ANS approaches on Aurora 2. (a) Test set A, (b) Test set B, and (c) Test set C.

As expected, by improving more advanced methods in the autocorrelation domain, i.e. DAS, SPFH and ANS using our proposed methods, better results were obtained in comparison to somewhat more basic autocorrelation-based methods, i.e. RAS and AMFCC. As it is clear, the combination of ANSS and overestimation with energy and cepstral mean and variance normalization (EMVN), overcame all other proposed methods in average overall performance on all the three test sets of AURORA 2. It is also worth mentioning that this performance is obtained with simple and low complexity computations, while ETSI-XAFE is a complicated algorithm with large computational overhead. Also, it is worth mentioning that, as will be shown in the appendix, the strongest advantage of the proposed methods over the ETSI-XAFE is at very low SNRs (-5dB in this case), which is not included in the figures reported in Table 2.

6. CONCLUSIONS

In this paper, we have raised the issue of using autocorrelation-based noise estimation and subtraction, taking into account the cross correlation term error. Two different methods were introduced for the insertion of the cross correlation term into the estimation of clean speech autocorrelation sequence, namely Kernel and ANSS. The Kernel method inserts the cross correlation term using a kernel function whereas ANSS considers the cross correlation term by averaging on a few frames. Both approaches were tested on Aurora 2 task and proved to be useful

in further improving the ANS results. Also, the overestimation parameter, as an important parameter where autocorrelation sequence estimation is concerned, was taken into account.

Practical experiments indicated that even better recognition performance could be expected when the overestimation parameter was introduced to ANS, Kernel, and ANSS methods. According to these results, although all the methods performed better when implemented in conjunction with the overestimation parameter, ANSS with overestimation parameter (ANSS+OEP) performed the best among them and its combination with energy and cepstral mean and variance normalization performed even better than the ETSI-XAFE. Altogether, a major result is that the features extracted from the autocorrelation sequence of the speech signal perform rather well in the presence of noise and the so-called mismatch conditions.

Table 2. Comparison of Average Recognition Rates and Percentage of Improvement in Comparison to MFCC for Various Feature Types on Three Test Sets of Aurora 2 Task

Feature type	Recognition rate (%)			Percentage of improvement (%)			Overall Average
	Set A	Set B	Set C	Set A	Set B	Set C	
MFCC	61.13	55.57	66.68	—	—	—	61.13
AMFCC	63.41	57.67	69.72	5.87	4.73	9.12	63.60
RAS	66.77	60.94	71.81	14.51	12.0	15.4	66.51
DAS	70.90	65.57	77.17	25.14	22.5	31.4	71.21
SPFH	73.61	68.98	80.89	32.11	30.1	42.6	74.49
MFCC-SS	69.22	63.46	73.60	20.81	17.7	20.7	68.76
MVA	76.05	76.35	73.10	38.38	46.7	19.2	75.17
ANS	77.10	74.32	83.61	41.09	42.2	50.8	78.34
Kernel	78.90	75.88	84.53	45.72	45.7	53.5	79.77
ANSS	80.47	79.04	85.53	49.76	52.8	56.5	81.86
ANS+OEP	78.78	75.98	84.14	45.41	45.9	52.4	79.63
Kernel+OEP	80.05	77.86	85.40	48.68	50.1	28.0	81.10
ANSS+OEP	82.37	81.10	86.21	54.64	57.4	58.6	83.23
ANSS+OEP+ EMVN	84.81	86.63	87.97	60.92	69.9	63.9	86.47
ETSI-XAFE	86.56	85.19	83.49	65.42	66.6	50.4	85.08

REFERENCES

[1] Boll, S., (1979), "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No. 2, pp 113-120.

[2] Evans, N. W. D., Mason, J. S. D., Liu, W. M. & Fauve, B., (2006). "An assessment on the fundamental limitations of spectral subtraction". Proceedings of ICASSP.

[3] McGinn, D.P. & Johnson, D.H., (1989), "Estimation of all-pole model parameters from noise-corrupted sequence", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, No. 3, pp 433-436.

[4] Jalil, M. Butt & Malik, A., (2013), "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals", Proceeding of TAEECE, pp 208-212.

[5] Amado, R.-G. & Filho, J.-V., (2008), "Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes", Proceeding of ICALIP, pp 449 – 454.
[6] Bansal Dev, P.,A. & BalaJain, S., (2009), "Role of Spectral Peaks in Autocorrelation Domain for Robust Speech Recognition", Journal of Computing and Information Technology-CIT17, Vol. 3, pp 295–303.
[7] Mansour, D. & Juang, B.-H., (1989), "The short-time modified coherence representation and noisy speech recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, No. 6, pp 795-804.
[8] Hernando, J. & Nadeu, C., (1997), "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition, IEEE Transactions on Speech and Audio Processing", Vol. 5, No. 1, pp 80-84.
[9] Yuo K.-H. & Wang, H.-C., (1999), "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences", Speech Communication, Vol. 28, pp 13-24.
[10] Shannon, B. J. & Paliwal, K. K., (2006). "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition", Speech Communication, Vol. 48, pp 1458–1485.
[11] Farahani, G., Ahadi, S.M. & Homayounpour, M.M., (2007), "Features based on filtering and spectral peaks in autocorrelation domain for robust speech recognition", Computer Speech and Language, Vol. 21, pp 187-205.
[12] Farahani, G., Ahadi, S.M. & Homayounpour, M.M., (2006), "Use of spectral peaks in autocorrelation and group delay domains for robust speech recognition", Proceeding of ICASSP.
[13] Farahani, G., Ahadi, S.M. & Homayounpour, M.M., (2006), "Robust Feature Extraction based on Spectral Peaks of Group Delay and Autocorrelation Function and Phase Domain Analysis", Proceeding of ICSLP.
[14] Ikbal, S., Misra, H. & Bourlard, H., (2003), "Phase autocorrelation (PAC) derived robust speech features", Proceeding of ICASSP, pp 133-136.
[15] Hu, Y., Bhatnagar, M. & Loizou, P., (2001), "A Cross-correlation Technique for Enhancing Speech Corrupted with Correlated Noise", Proceeding of ICASSP.
[16] Farahani, G., Ahadi, S.M. & Homayounpour, M.M., (2006), "Robust feature extraction using spectral peaks of the filtered higher lag autocorrelation sequence of the speech signal", Proceeding of ISSPIT.
[17] Farahani, G., Ahadi, S.M. & Homayounpour, M.M., (2006), "Robust Feature Extraction of Speech via Noise Reduction in Autocorrelation Domain", Proceeding of IWMRCS.
[18] SPIB, (1995), SPIB noise data. Available from: <http://spib.rice.edu/spib/select_noise.html>.
[19] Lu, Y. & Loizou, P.C., (2008), "A geometric approach to spectral subtraction", Speech Communication, Vol. 50, pp 453–466.
[20] Onoe, K., Segi, H., Kobayakawa, T., Sato, S., Imai, T. & Ando, A., (2002), "Filter Bank Subtraction for Robust Speech Recognition", Proceeding of ICSLP.
[21] Kitaoka, N. & Nakagawa, S., (2002), "Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task", Proceeding of ICSLP, pp 477-480.
[22] Vasseghi, S.V., (2006), Advanced Digital Signal Processing and Noise Reduction, 3rd Edition, John Wiley & Sons Ltd.
[23] Kamath, S. D. & Loizou, P. C., (2002), "A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise", Proceeding of ICASSP.
[24] Farahani, G., Ahadi, S.M. & Homayounpour, M.M., (2007), "Improved Autocorrelation-based Noise Robust Speech Recognition Using Kernel-Based Cross Correlation and Overestimation Parameters", Proceeding of EUSIPCO.
[25] Hirsch, H.G. & Pearce, D., (2000), "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", Proceeding of ISCA ITRW ASR.
[26] HTK, (2002), The hidden Markov model toolkit available from: <http://htk.eng.cam.ac.uk>.
[27] Ahadi, S. M., Sheikhzadeh, H., Brennan, R. L. & Freeman, G. H., (2004), "An Energy Scheme for Improved Robustness in Speech Recognition", Proceeding of ICSLP.
[28] Chen, C.-P, Bilmes, J. & Kirchhoff, K., (2002), "Low-Resource Noise-Robust Feature Post-Processing on Aurora 2.0", Proceeding of ICSLP, pp 2445-2448.

[29] ETSI-XAFE, (2003), ETSI Standard, Extended advanced front-end feature extraction algorithm - ETSI ES 202 212 V1.1.1.

AUTHOR

Gholamreza Farahani received his BSc degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1998 and MSc and PhD degrees in electrical engineering from Amirkabir University of Technology (Polytechnic), Tehran, Iran in 2000 and 2006 respectively. Currently, he is an assistant professor in the Institute of Electrical and Information Technology, Iranian Research Organization for Science and Technology (IROST), Iran. His research interest is signal processing especially speech processing