Farahani *EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:13
DOI 10.1186/s13636-017-0110-8

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Autocorrelation-based noise subtraction method with smoothing, overestimation, energy, and cepstral mean and variance normalization for noisy speech recognition

Gholamreza Farahani

Abstract

Autocorrelation domain is a proper domain for clean speech signal and noise separation. In this paper, a method is proposed to decrease effects of noise on the clean speech signal, autocorrelation-based noise subtraction (ANS). Then to deal with the error introduced by assumption that noise and clean speech signal are uncorrelated, two methods are proposed. Also to improve recognition rate of speech recognition system, overestimation parameter is used. Finally, with the addition of energy and cepstral mean and variance normalization to features of speech, recognition rate has improved considerably in comparison to standard features and other correlation-based methods. The proposed methods are tested on the Aurora 2 database. Between different proposed methods, autocorrelation-based noise subtraction method with smoothing, overestimation, energy, and cepstral mean and variance normalization (ANSSOEMV) method has a best recognition rate improvement in average than MFCC features which is 64.91% on the Aurora 2 database.

Keywords: Autocorrelation-based methods, Noise subtraction, Robustness, Recognition rate, Speech recognition

1 Introduction

The accuracy of speech recognition systems will degrade severely when the systems are operated in adverse acoustical environments. 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. In recent years, many approaches have been developed to address the problem of robust speech recognition.

These methods can be very roughly classified into model-based and feature-based, which in this paper our concentration is on the 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 (MFCCs) that have shown good performance in clean-train/clean-test conditions, but deteriorated performance in the cases of mismatch.

In Meyer et al. [28], the condition of mismatches is created and with behavioral experiment has reviewed the effects of such acoustic disturbances on speech perception in conditions approaching ecologically valid contexts.

A very well-known and widely used enhancement method that deals with the signal spectrum is spectral

Correspondence: farahani.gh@irost.org
Electrical and Information Technology Institute, Iranian Research Organization for Science and Technology (IROST), Sh. Ehsani Rad St., Enqelab St., Parsa Sq., Ahmadabad Mostoufi Rd., Azadegan Highway, Tehran 3353136846, Iran, Islamic Republic of

subtraction (SS) [4]. 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 [9], 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 [27]. 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 Jalil et al. [22], 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 are introduced.

Pitch detection algorithms (PDAs) for simple audio signals based on zero-cross rate (ZCR) and autocorrelation function (ACF) in Amado and Filho [2] are presented also in Muhammad [29], with use of autocorrelation function noise-robust pitch detection is performed, and experimental results have shown the superiority of proposed method over other methods.

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, is among the most successful methods; we can name differentiated relative autocorrelation sequence spectrum (DRASS) [3], short-time modified coherence (smc) [26], one-sided autocorrelation LPC (OSA LPC) [17], relative autocorrelation sequence (RAS) [36], autocorrelation mel-frequency cepstral coefficients (AMFCC) [33], and differentiation of autocorrelation sequence (DAS) [14]. 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 [10, 11].

In DRASS, autocorrelation will be calculated by biased estimator after frame blocking and pre-emphasis. Then after filtering, FFT will be calculated and absolute amplitude of differentiated FFT square amplitude will be used for mel-scale frequency bank. Finally, log of coefficients and their cepstrum will be used 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 OSA LPC, 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 are found as the final coefficients. Among the 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 [21] and autocorrelation peaks and phase features (APP) [11].

Noise-robust feature extraction method for speech recognition using the robust perceptual minimum variance distortionless response (MVDR) spectrum of temporally filtered autocorrelation sequence is proposed in Seyedin et al. [32] which has improved speech recognition rate.

New set of (perceptual linear predictive) PLP vector is autocorrelation domain proposed in Rahali et al. [31] and tested in various noisy conditions, and significant improvement is obtained in comparison to traditional feature extraction techniques.

In DRHOASS, differential of relative higher order of autocorrelation sequence spectrum will calculate which results show these new features more robust than traditional MFCC features in additive noise conditions [7].

Selecting the number of feature coefficients is important for speech recognition accuracy. Therefore, in this paper Fisher-Markov selector is evaluated to identify those features that are most useful in speech recognition [6].

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 of autocorrelation function. Section 3 will describe some autocorrelation-based approaches that are related to our proposed algorithms. Section 4 is dedicated to discussion on our proposed method for speech recognition enhancement in autocorrelation domain. In Section 5, some implementation issues regarding the proposed methods will be addressed. Section 6 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 7 will conclude the paper.

2 Theory of autocorrelation function

In this section, we will describe the theory of autocorrelation function for speech signal. This section will give us an appropriate insight to the advantages and disadvantages of using autocorrelation function in robust feature extraction.

2.1 Formulation of clean speech signal, noise, and noisy speech signal in autocorrelation domain

First relationship between the autocorrelation sequences of clean, noise and noisy signals will be explain. If we assume $v(m,n)$ to be the additive noise and $x(m,n)$ to be clean speech signal, then noisy speech signal, $y(m,n)$, could be written as

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

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 from noisy speech signal, the channel distortion effects are not considered in the equations. Generally, clean speech signal and noise will consider uncorrelated, therefore, if $x(m,n)$ and $v(m,n)$ are considered uncorrelated, then the autocorrelation of the noisy speech signal can be written as

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

where $r_{yy}(m,k)$, $r_{xx}(m,k)$, and $r_{vv}(m,k)$ 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, that clean speech signal and noise have correlation Eq. (2) should be written as

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

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.

With an assumption that noise autocorrelation sequence is relatively constant across frames, we can find an estimate of $r_{vv}(m,k)$ using the non-speech sections of an utterance specified, for example, by a voice activity detector (VAD) or by the initial normally non-speech periods and denote it as $\hat{r}_{vv}(k)$. Then Eq. (4) can be written as

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

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 [20].

3 Review of autocorrelation-based methods for robust feature extraction in noisy condition

Until now, several autocorrelation-based methods have been proposed, where usually the speech signal and noise were considered uncorrelated. In order to get some insight on how autocorrelation properties may be used to achieve robustness, some of these methods are described.

3.1 Relative autocorrelation sequence (RAS) method

As explained in Yuo and Wang [36], this method assumed the noise as stationary and uncorrelated to the clean speech signal. Therefore, the relationship between the autocorrelations of noisy and clean signals and noise could be written as

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

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

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

The right side of Eq. (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 (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 6 for RAS performance in clean speech conditions).

3.2 Autocorrelation mel-frequency cepstral coefficients (AMFCC) method

In this approach [33], the MFCC coefficients were extracted from the noisy speech 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 speech signal for many noise types, including those with least correlations among frames. The lag threshold value used was 3 ms and was set by finding the first valley in the absolute autocorrelation function found over TIMIT speech frames.

As reported in Shannon and Paliwal [33], 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.

3.3 Differentiation of autocorrelation sequence (DAS) method

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 [14].

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

$$\mathrm{Diff}_Y(k) \approx \sum_{l=-Q}^{P} a_l.Y(k+l), \quad 0 \le k \le K-1 \tag{8}$$

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) [12]. The differentiation mentioned in Eq. (8) can be carried out in several ways, as discussed in Farahani et al. [12]. The simple difference had shown the best results and therefore was used in Farahani et al. [14], i.e.,

$$\mathrm{Diff}_Y(k) = Y(k) - Y(k+1) \tag{9}$$

The procedure of feature extraction was carried out after high-pass filtering (as in Eq. (7)) and peak extraction (as in Eq. (9)). As explained earlier for RAS, this filtering can suppress the effect of slowly varying noises and can 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.

3.4 Spectral peaks of filtered higher-lag autocorrelation sequence (SPFH) method

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 [13].

In SPFH, after frame blocking and pre-emphasis of the noisy signal, the autocorrelation sequence of the frame signal was obtained as in Eq. (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 Eq. (7). Then, hamming windowing and short-time fourier transform were carried out, and the differential power spectrum of the filtered signal was found using Eq. (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 filterbank 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 Eq. (7), and preserving spectral peaks, as in equation (9), i.e. similar to DAS.

4 Proposed methods for speech recognition enhancement in noisy conditions at autocorrelation domain

4.1 Autocorrelation-based noise subtraction (ANS) method

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 in Section 3. 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 clean speech signal. As an alternative to such methods, we follow a newer approach. In ANS method, 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 [13]. An instant advantage of ANS method is that there is no need to deal with phase issue.

In Farahani et al. [13], 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

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

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

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

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

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.

4.2 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 as Eq. (6). If we square both sides of autocorrelation magnitude in Eq. (6), then each frame will be as follows:

$$\left|r_{\mathrm{yy}}(m)\right|^2=|r_{\mathrm{xx}}(m)|^2+|r_{\mathrm{vv}}|^2+2.|r_{\mathrm{xx}}(m)|.|r_{\mathrm{vv}}|.\cos\theta(m) \quad 0 \le m \le M-1 \quad (12)$$

where $r_{\mathrm{yy}}(m)$, $r_{\mathrm{xx}}(m)$ and r_{vv} are the short-time one-sided autocorrelation vectors of the noisy speech, clean speech, and noise in each frame index m, respectively, and according to the autocorrelation definition) dot products of two vectors), $\theta(m)$ is the instantaneous phase difference between clean speech signal autocorrelation, $r_{\mathrm{xx}}(m)$, and noise autocorrelation, r_{vv} or in other words, $\theta(m)$ is the angle between autocorrelation of $r_{\mathrm{xx}}(m)$ and r_{vv} vectors, $|r_{\mathrm{yy}}(m)|$, $|r_{\mathrm{xx}}(m)|$, and $|r_{\mathrm{vv}}|$ are noisy speech, clean speech, and noise autocorrelation amplitude, respectively. From Eq. (12), we will have [30]

$$\begin{aligned}|r_{\mathrm{xx}}(m)|^2&=\left|r_{\mathrm{yy}}(m)\right|^2-|r_{\mathrm{vv}}|^2.(1+2r(m).\cos\theta(m))\\&=\left|r_{\mathrm{yy}}(m)\right|^2-M(r(m),\theta(m)).|r_{\mathrm{vv}}|^2\end{aligned} \quad (13)$$

Where

$$r(m)=\frac{|r_{\mathrm{xx}}(m)|}{|r_{\mathrm{vv}}|}$$

$$M(r(m),\theta(m))=1+2r(m).\cos\theta(m) \quad (14)$$

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

The variation of the kernel function $M(r(m), \theta(m))$ in a frame is drawn in Fig. 1. We normalized $|r_{vv}|$ between 0 ~ 1 and named it $|d|$. Also $\theta(m)$ changes between $-\pi$ and $+\pi$ with clean speech amplitude equal to 1.

As it is clear from Fig. 1, when the noise autocorrelation amplitude $|r_{vv}|$ is large, changes in $\theta(m)$ result in large changes in $M(r(m), \theta(m))$.

In the following equation, we have the noise autocorrelation component of each frame as [30]

$$\begin{aligned} z(m) &= |r_{yy}(m)| - |r_{xx}(m)| \\ &= \sqrt{|r_{xx}(m)|^2 + |r_{vv}|^2(1 + 2r(m).\cos\theta(m))} - |r_{xx}(m)| \\ &= |r_{vv}|\left(\sqrt{r^2(m) + 2r(m).\cos\theta(m) + 1} - r(m)\right) \end{aligned} \tag{15}$$

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

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

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

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

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

$$z(m) = |r_{vv}|.\gamma(r(m)) \tag{18}$$

and the clean speech signal autocorrelation amplitude in each frame is estimated by

$$r_{xx}(m) = r_{yy}(m) - z(m) \tag{19}$$

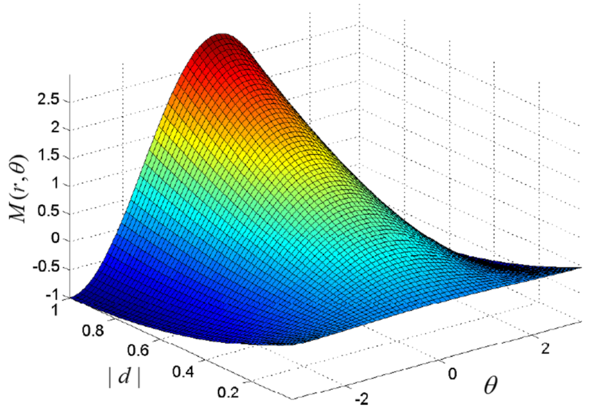


Fig. 1 Variation of $M(r(m), \theta(m))$ versus $|d|$ and $\theta(m)$

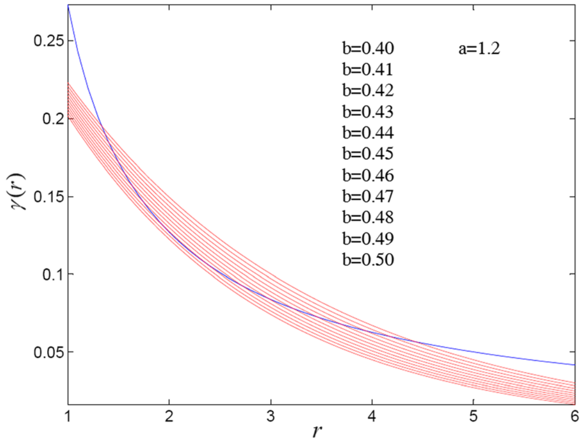


Fig. 2 Function $\gamma(r)$. The bunch of curves represents the estimation of the function according to different values of b. The other curve represents the real values of $\gamma(r)$

For the sake of simplicity, according to Fig. 2, we change the function $\gamma(r(m))$ 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

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

where a was set to 1.2 and b to 0.45 in our experiments.

Therefore, in our implementations, we have used (19) instead of (11). We named this method as kernel method.

b=.40 a=1.2

b=.41

b=.42

b=.43

b=.44

b=.45

b=.46

b=.47

b=.48

b=.49

b=.50

4.3 Consideration of cross-correlation term in noisy speech recognition

Figure 3 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 [34], as well as the sum of autocorrelation sequences of speech and noise. One should expect the clean speech signal and noise, in most circumstances, to be completely uncorrelated. However, in this case, according to Fig. 3, the autocorrelation sequence of the noisy speech signal 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 clean speech signal autocorrelation, one needs to consider some correlation among clean 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 Eq. (2) is only possible when an infinitely long signal is considered in the analysis [25]. We have used the two following approaches in order to consider the cross-correlation term in autocorrelation calculations:

4.3.1 Autocorrelation averaging method

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 Kitaoka and Nakagawa [24], 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

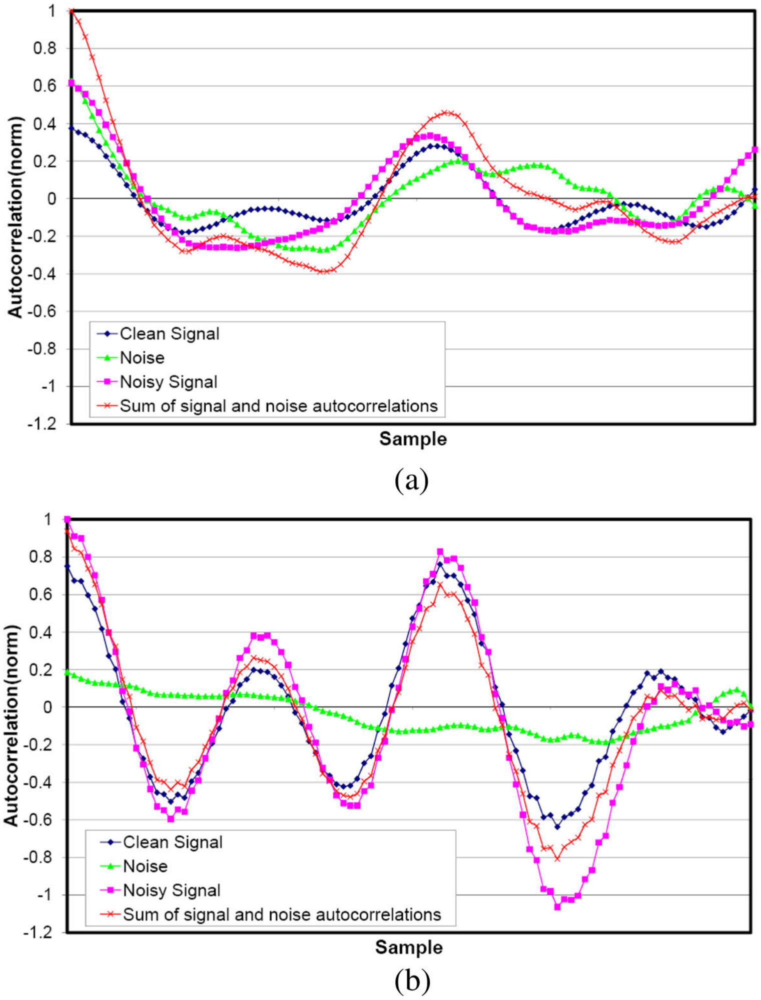


Fig. 3 Sample autocorrelation sequences of the clean speech, noisy speech, and noise as well as sum of the autocorrelations of clean speech signal and noise with **a** babble noise and **b** factory noise, with an SNR of 10 dB

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

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

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.

By replacing $r_{\mathrm{yy}}(m-i,k)$ in Eq. (21) with the value found in Eq. (4) we have

$$\begin{aligned}\overline{r_{\mathrm{yy}}(m,k)} &= \sum_{i=0}^{T-1} b_i . r_{\mathrm{yy}}(m-i,k) \qquad (22)\\ &= \sum_{i=0}^{T-1} b_i . r_{\mathrm{xx}}(m-i,k) + \sum_{i=0}^{T-1} b_i . r_{\mathrm{vv}}(m-i,k)\\ &+E\left\{\sum_{i=0}^{T-1} b_i . x(m-i,k).v^*(m-i,k)\right\}\\ &+E\left\{\sum_{i=0}^{T-1} b_i . x^*(m-i,k).v(m-i,k)\right\}\end{aligned}$$

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

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

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

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

and

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

Setting the value of the parameters T and b_i will be discussed in the parameter settings section (Section 5.3), and we will see that with the values used for T, the above mentioned assumption holds.

It was shown in Kitaoka and Nakagawa [24] that if the phase differences between the clean speech and noise in successive frames are assumed to be uncorrelated, the PDFs of the summation terms in Eqs. (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 Eq. (22) would be rewritten as

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

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

$$\begin{aligned}\hat{r}_{\mathrm{xx}}(m,k) = \overline{r_{\mathrm{yy}}(m,k)} - \hat{r}_{\mathrm{vv}}(m,k) \approx r_{\mathrm{xx}}(m,k)\\ + r_{\mathrm{vv}}(m,k) - \hat{r}_{\mathrm{vv}}(m,k)\end{aligned} \tag{28}$$

Therefore, if we estimate the autocorrelation sequence of noise, $\hat{r}_{\mathrm{vv}}(m,k)$ 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 6), this effect is negligible.

We will call this approach autocorrelation-based noise subtraction with smoothing (ANSS). Details of the setting of the length of averaging window in this approach will be discussed in the parameter setting section (5.3).

4.4 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 (temporal 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 [35]. This non-linear 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 0 and 10 dB 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.

5 Subjects in implementation of proposed algorithms

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

5.1 Considering cross-correlation term in autocorrelation domain

To consider the cross-correlation term, we have implemented a method, as discussed in Subsections 4.3.1.

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 Subsection 5.3)
(e) Autocorrelation averaging calculation (see Subsection 4.3.1)

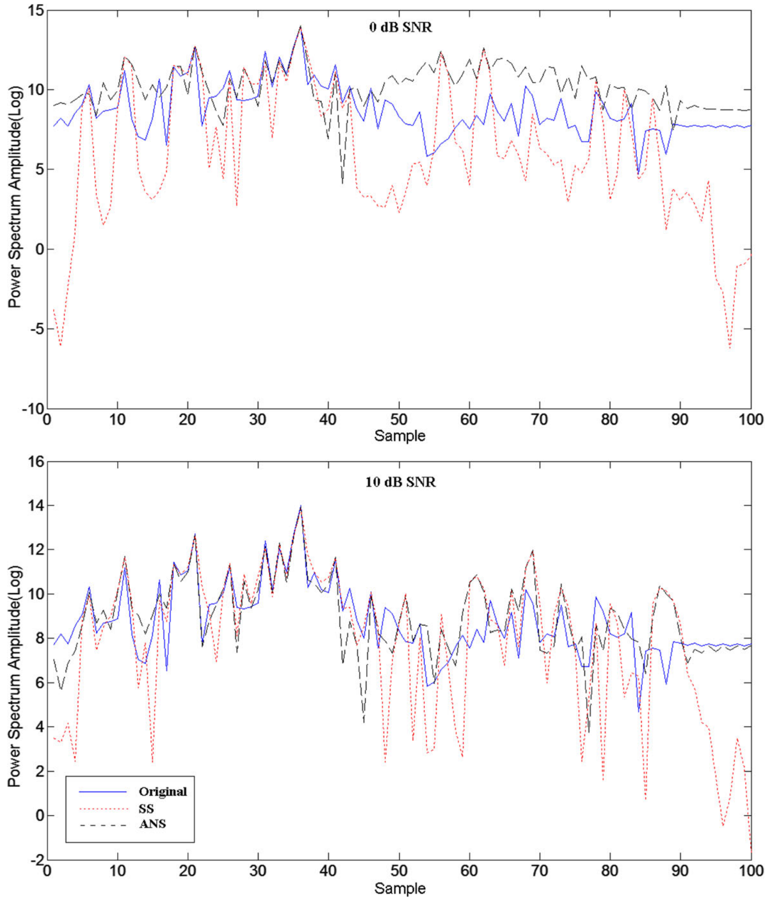


Fig. 4 Log power spectrum of a speech frame of ‘FAK_1B.08’ utterance from test set A of Aurora 2 task contaminated with subway noise at 0 and 10 dB SNRs in logarithmic scale

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					
	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
SS	332	378	488	653	982	1500
ANS	22.4	28.1	39.3	55.9	87.9	142

(f) Inserting cross-correlation term in the estimation of the autocorrelation of the clean speech signal
(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 previous step
(j) Calculation of the feature vectors including 12 cepstral and a log-energy parameter and their first and second order dynamic parameters

In this algorithm, 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 6.

5.2 Considering overestimation parameter in autocorrelation methods

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 Fig. 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

$$r_{\mathrm{yy}}(m,k) = r_{\mathrm{xx}}(m,k) + \alpha.\hat{r}_{\mathrm{vv}}(k), \tag{29}$$

where $\alpha \geq 1$ is the overestimation parameter. Note that when $\alpha = 1$, Eq. (29) is identical to the equation used for ANS.

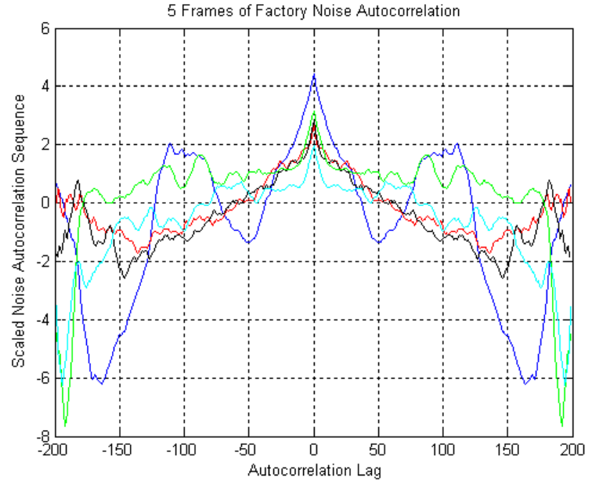


Fig. 5 Autocorrelation sequences for five consecutive frames of factory noise

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 Eq. (29). Various values of α were tested to get the best result on the Aurora 2 task.

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 (5.3). 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 Fig. 6, according to changes observed in system recognition performance in practice [13]. 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 4.3, together with the overestimation parameter,

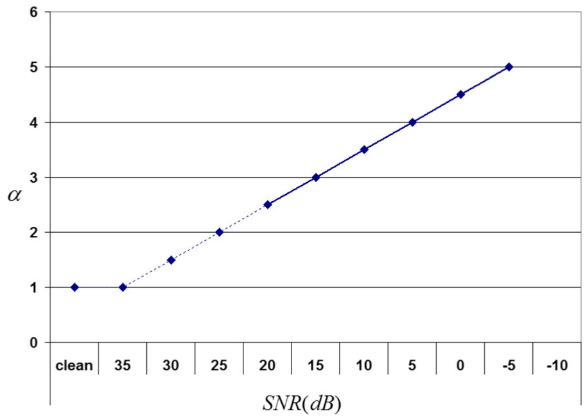


Fig. 6 Change in the parameter *α* with SNR on Aurora 2 task

the following relationship for clean speech signal estimation will result in

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

Meanwhile, considering the cross-correlation term as in Section 4.3.1, together with the overestimation parameter, we will have the following equation, which gives an approximate value of the speech signal.

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

5.3 Parameter settings

In our implementation of RAS, the length of the filter was set to $L = 2$ according to Yuo and Wang [36]. 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 Shannon and Paliwal [33]. The same duration was also used for SPFH implementation [12].

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 Farahani et al. [15], 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 Eq. (21) to Eq. (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 three 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.

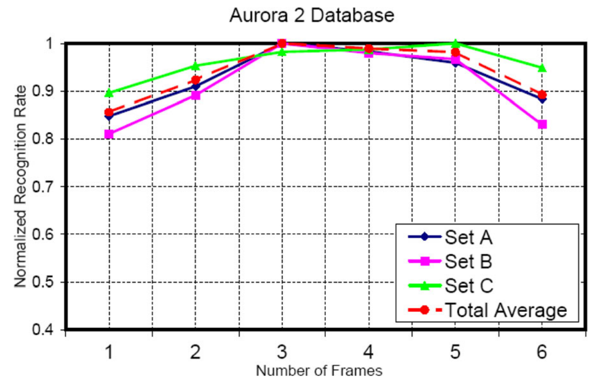


Fig. 7 Normalized recognition rates for test sets of Aurora 2 task versus the number of frames used in noise autocorrelation sequence averaging and their grand average

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 Eq. (32).

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

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 Fig. 6. The parameter setting for Fisher-Markov selector is carried out similar to Hegde et al. [16].

6 Experiments and comparison of methods

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

6.1 Database

The experiments were carried out on Aurora 2 task [18]. The features in this case were computed using 25 ms frames with 10 ms 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. Also with use of Fisher-Markov selector, comparison of different number of MFCC coefficients will carry out [6]. All model creation, training, and tests in all our experiments have been carried out using the standard hidden Markov model toolkit [19] with 16 states and 3 mixture components per state. The HMMs were trained in clean condition, i.e., with clean training data.

6.2 Number of MFCC features using Fisher-Markov selector

According to Hegde et al. [16], 8 MFCC coefficients, with the use of Fisher's ratio technique, could have better classification accuracy than other number of coefficients 3 to 12 MFCCs for 5 vowels in *Kannada* language. In this paper, the results of MFCC for different number of coefficients with Fisher-Markov selector are evaluated on the Aurora 2.0 database. The average recognition results are shown in Table 2.

Figure 8 shows the effects of the number of MFCC on the recognition rate for each set of Aurora 2 database and overall average of it with Fisher-Markov selector. As it is clear from Table 2 and Fig. 8, the best recognition

Table 2 Comparison of overall average recognition rates of different number of MFCC features with Fisher-Markov selector on three test sets of Aurora 2 task

Number of MFCC coefficients	Recognition rate (%)			Overall average of recognition rate (%)
	Set A	Set B	Set C	
3 MFCC + log energy + their first and second order derivatives	57.38	44.95	51.29	51.21
4 MFCC + log energy + their first and second order derivatives	58.48	46.25	52.65	52.46
5 MFCC + log energy + their first and second order derivatives	60.41	48.16	54.46	54.34
6 MFCC + log energy + their first and second order derivatives	60.27	48.04	54.34	54.22
7 MFCC + log energy + their first and second order derivatives	62.49	50.13	56.45	56.36
8 MFCC + log energy + their first and second order derivatives	64.87	52.98	58.31	58.72
9 MFCC + log energy + their first and second order derivatives	64.23	52.01	57.48	57.91
10 MFCC + log energy + their first and second order derivatives	65.24	53.18	59.12	59.18
11 MFCC + log energy + their first and second order derivatives	64.56	52.63	57.98	58.39
12 MFCC + log energy + their first and second order derivatives	65.96	54.78	60.32	60.35

rate of MFCC with Fisher-Markov selector will obtain 12 MFCC. Although the Fisher-Markov selector in comparison to MFCC reduces the recognition rate slightly, but it will reduce the times of MFCC feature calculations considerably. Therefore, for our comparison with other autocorrelation-based methods, because in this paper our target is maximization of recognition rate, we used 12 MFCCs and a log-energy parameter, together with their first and second order derivatives.

6.3 Implementation results using cross-correlation terms

The setting of our parameters is as described in 5.3. Figure 9 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 4. 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.

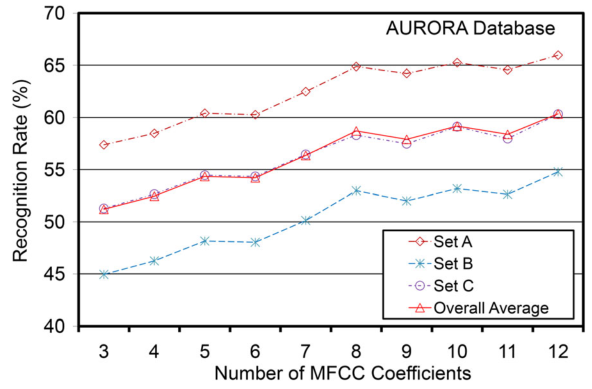
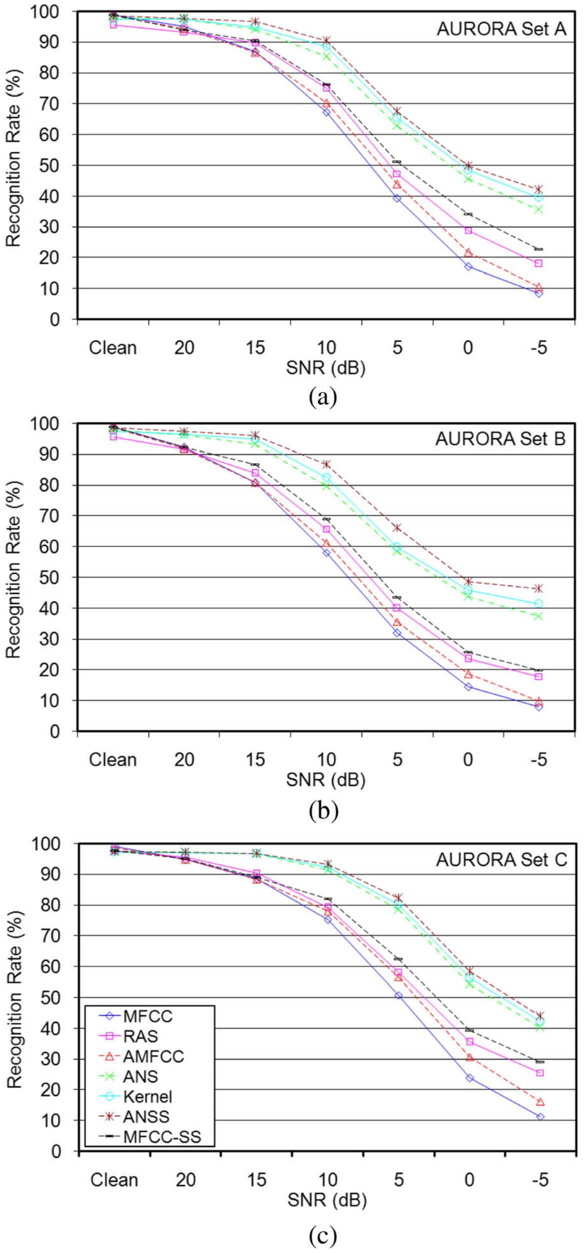


Fig. 8 Variation of recognition rate with Fisher-Markov selector for different number of MFCC on the Aurora 2 database

This is indicative of the effectiveness of inserting the cross-correlation parameter into the autocorrelation calculation of noisy speech signal.

6.4 Implementation results of applying overestimation parameter to the proposed methods

The results of including the overestimation parameter α into clean speech autocorrelation estimation procedure

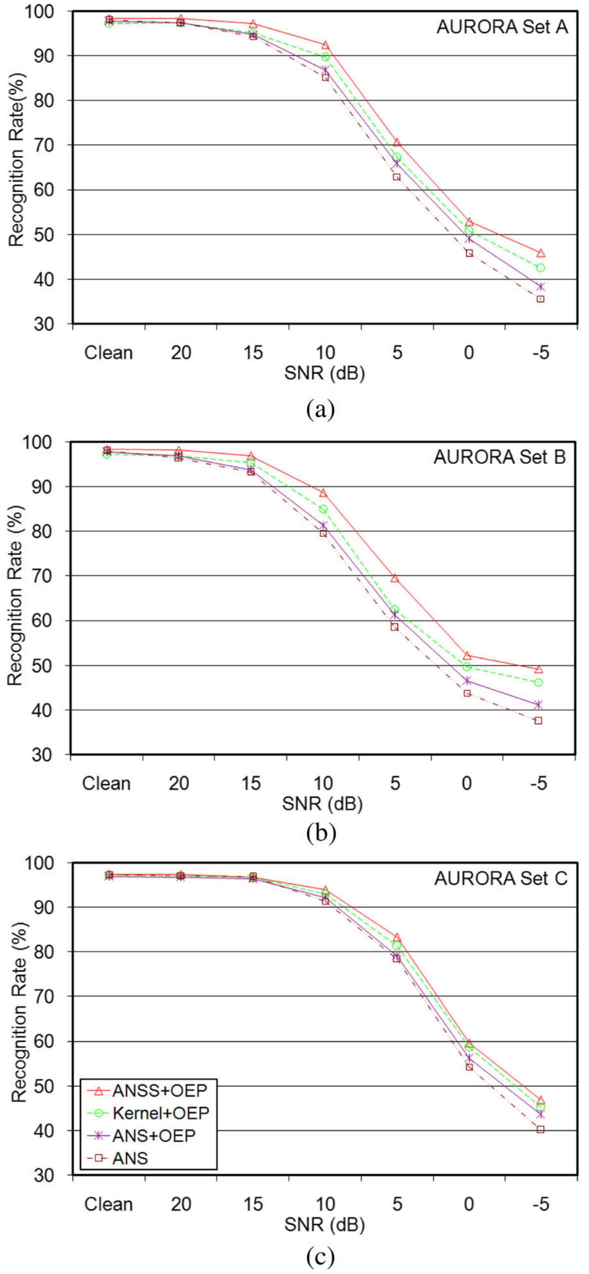


(a)

(b)

(c)

Fig. 9 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. **c** Test set C



(a)

(b)

(c)

Fig. 10 Average recognition rates of ANSS + OEP, kernel + OEP, ANS + OEP, and ANS approaches on Aurora 2. **a** Test set A. **b** Test set B. **c** Test set C

will be reported here. Figure 10 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.

6.5 Comparison the results of different methods in autocorrelation domain

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 Ahadi et al. [1], 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 3 shows the average recognition rates of all these methods on the Aurora 2 task. As usual in Aurora 2 result calculations, the −5 dB 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) [5]. 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 [8] on the Aurora 2 corpus are reported. This is a standard front-end which uses sophisticated enhancement approaches to improve the quality of the extracted features. Further details about its performance are reported in the Appendix.

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), named ANSSOEMV (autocorrelation-based noise subtraction method with smoothing, overestimation, energy, and cepstral mean and variance normalization), overcame all other proposed methods in average overall performance on all the three test sets of Aurora 2 which is 64.91% in average than MFCC features. 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 (−5 dB in this case), which is not included in the figures reported in Table 3.

7 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 were proved to be useful in further improving the ANS results. Also, the overestimation parameter, as

Table 3 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	Overall average improvement (%)
	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	6.57
RAS	66.77	60.94	71.81	14.51	12.09	15.40	66.51	14.00
DAS	70.90	65.57	77.17	25.14	22.51	31.48	71.21	26.37
SPFH	73.61	68.98	80.89	32.11	30.18	42.65	74.49	34.98
MFCC-SS	69.22	63.46	73.60	20.81	17.76	20.77	68.76	19.78
MVA	76.05	76.35	73.10	38.38	46.77	19.27	75.17	34.81
ANS	77.10	74.32	83.61	41.09	42.20	50.81	78.34	44.70
Kernel	78.90	75.88	84.53	45.72	45.71	53.57	79.77	48.33
ANSS	80.47	79.04	85.53	49.76	52.82	56.57	81.68	53.05
ANS + OEP	78.78	75.98	84.14	45.41	45.94	52.40	79.63	47.92
Kernel + OEP	80.05	77.86	85.40	48.68	50.17	56.18	81.10	51.68
ANSS + OEP	82.37	81.10	86.21	54.64	57.46	58.61	83.23	56.91
ANSSOEMV	84.81	86.63	87.97	60.92	69.91	63.90	86.47	64.91
ETSI-XAFE	86.56	85.19	83.49	65.42	66.67	50.45	85.08	60.85

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 (ANSSOEMV) 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.

8 Endnote

[1] A more detailed discussion on this issue can be found in [30].

9 Appendix

9.1 Comparison with ETSI extended advanced front-end

In order to be able to compare the performance of our robust speech recognition approach with a standardized front-end, we implemented ETSI extended advanced front-end [8] on the Aurora 2 corpus. This is a standard front-end with a sophisticated enhancement scheme that tries to improve the quality of speech signal before extracting features. The ETSI-XAFE was implemented using the tools provided by ETSI.

The results of our ANSS + OEP method after the application of energy, mean, and variance normalization, abbreviated as ANSSOEMV, and the mean subtraction, variance normalization, and ARMA filtering (MVA) approach are also reported here. Figure 11 depicts the recognition results on Aurora 2 task.

As it is clear, the application of overestimation with mean and variance normalization of the cepstral parameters has led to improvements in the system recognition performance, which in comparison to ETSI-XAFE has a better recognition rate in most of the cases. These results also are indicative of the effect of mean and variance normalization of cepstral coefficients in improving autocorrelation-based methods for noise reduction.

One interesting point is the noticeable difference between the performance of our approach and that of the other two in very low SNR (−5 dB) in all three test sets. The experimental results have shown that all ANS-based methods perform better than ETSI-XAFE at this SNR, indicating that the methodology used in autocorrelation-based approaches is performing very well in very low SNRs, in comparison to that of ETSI-XAFE. In fact, it is clear from both Figs. 9 and 10 that all the autocorrelation-based methods as well as spectral subtraction follow an almost similar trend of performance change in low SNRs, while this is quite different for ETSI-XAFE and MVA. This could be attributed to the noise estimation procedure followed in these methods that loses its performance with decreasing SNR gradually, while the other two methods seem to be quite sensitive to high levels of noise.

It is also worth mentioning that our approach was completed about 30% faster than ETSI-XAFE on the mentioned task when run on a Pentium 4-based computer.

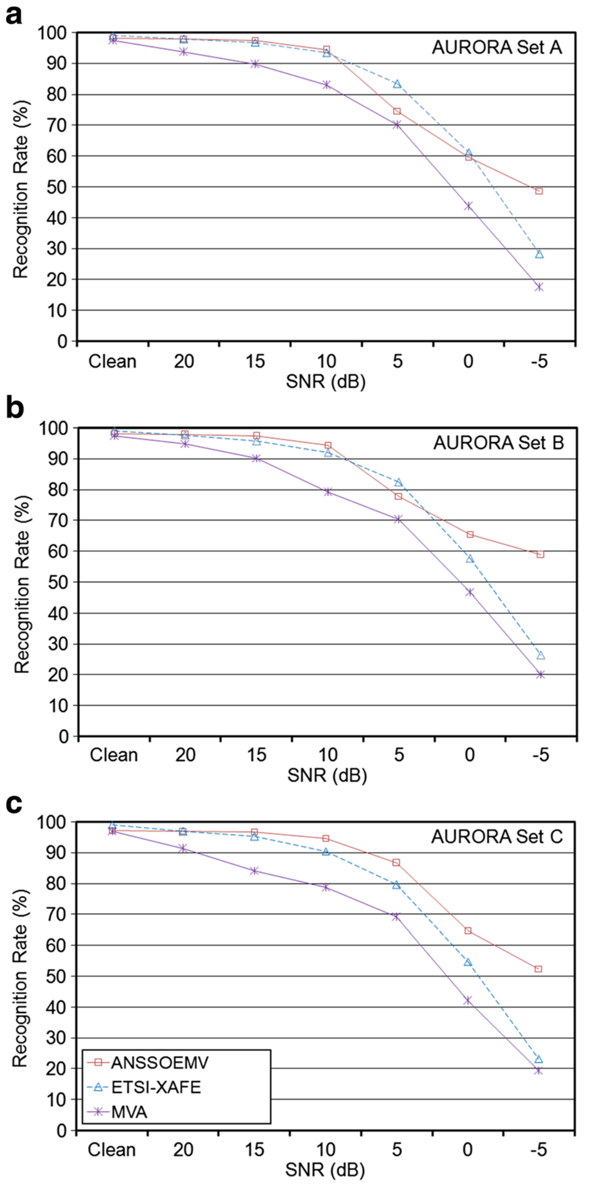


Fig. 11 Average recognition rates of ANSSOEMV, ETSI-XAFE, and MVA approaches on Aurora 2. **a** Test set A, **b** Test set B. **c** Test set C

Author's information

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.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 10 February 2017 Accepted: 4 June 2017
Published online: 21 June 2017

References

1. SM Ahadi, H Sheikhzadeh, RL Brennan, GH Freeman, *An energy scheme for improved robustness in speech recognition*. Proceeding of ICSLP, 2004
2. R-G Amado, J-V Filho, *Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes*. Proceeding of ICALIP, 2008, pp. 449–454
3. P Bansal, A Dev, S BalaJain, Role of spectral peaks in autocorrelation domain for robust speech recognition. Journal of Computing and Information Technology-CIT17 **3**, 295–303 (2009)
4. S Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Transaction on Acoustics, Speech and Signal Processing ASSP **27**, 113–120 (1979)
5. C-P Chen, J Bilmes, K Kirchhoff, *Low-resource noise-robust feature post-processing on Aurora 2.0*. Proceeding of ICSLP, 2002, pp. 2445–2448
6. Q Chemg, H Zhou, J Cheng, The Fisher-Markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. IEEE Transaction on Pattern Analysis and Machine Intelligence **33**, 6 (2011)
7. A Dev, B Poonam, Robust features for noisy speech recognition using MFCC computation from magnitude spectrum of higher order autocorrelation coefficients. International Journal of Computer Applications **10**, 8 (2010)
8. ETSI-XAFE, *ETSI Standard, Extended Advanced Front-end Feature Extraction Algorithm - ETSI ES 202 212 V1.1.1*, 2003
9. NWD Evans, JSD Mason, WM Liu, B Fauve, *An assessment on the fundamental limitations of spectral subtraction*. Proceedings of ICASSP, 2006
10. G Farahani, SM Ahadi, MM Homayounpour, *Use of spectral peaks in autocorrelation and group delay domains for robust speech recognition*. Proceeding of ICASSP, 2006
11. G Farahani, SM Ahadi, MM Homayounpour, *Robust feature extraction based on spectral peaks of group delay and autocorrelation function and phase domain analysis*. Proceeding of ICSLP, 2006
12. G Farahani, SM Ahadi, MM Homayounpour, *Robust feature extraction using spectral peaks of the filtered higher lag autocorrelation sequence of the speech signal*. Proceeding of ISSPIT, 2006
13. G Farahani, SM Ahadi, MM Homayounpour, *Robust feature extraction of speech via noise reduction in autocorrelation domain*. Proceeding of IWMRCS, 2006
14. G Farahani, SM Ahadi, MM Homayounpour, Features based on filtering and spectral peaks in autocorrelation domain for robust speech recognition. Computer Speech and Language **21**, 187–205 (2007)
15. G Farahani, SM Ahadi, MM Homayounpour, *Improved autocorrelation-based noise robust speech recognition using kernel-based cross correlation and overestimation parameters*. Proceeding of EUSIPCO, 2007
16. S Hegde, KK Achary, S Shetty, Feature selection using Fisher's ratio technique for automatic speech recognition. International Journal on Cybemetics & Informatics (IJCI) **4**, 2 (2015)
17. J Hernando, C Nadeu, Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. IEEE Transactions on Speech and Audio Processing **5**(1), 80–84 (1997)
18. HG Hirsch, D Pearce, *The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions*. Proceeding of ISCA ITRW ASR, 2000
19. HTK, *The hidden Markov model toolkit*, 2002. available from: http://htk.eng.cam.ac.uk
20. Y Hu, M Bhatnagar, P Loizou, *A cross-correlation technique for enhancing speech corrupted with correlated noise*. Proceeding of ICASSP, 2001
21. S Ikbal, H Misra, H Bourlard, *Phase autocorrelation (PAC) derived robust speech features*. Proceeding of ICASSP, 2003, pp. 133–136
22. M Jalil, F-A Butt, A Malik, *Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals*. Proceeding of TAEECE, 2013, pp. 208–212
23. SD Kamath, PC Loizou, *A multi-band spectral subtraction method for enhancing speech corrupted by colored noise*. Proceeding of ICASSP, 2002
24. N Kitaoka, S Nakagawa, *Evaluation of spectral subtraction with smoothing of time direction on the Aurora 2 task*. Proceeding of ICSLP, 2002, pp. 477–480
25. Y Lu, PC Loizou, A geometric approach to spectral subtraction. Speech Communication **50**, 453–466 (2008)
26. D Mansour, B-H Juang, The short-time modified coherence representation and noisy speech recognition, IEEE transactions on acoustics. Speech and Signal Processing **37**(6), 795–804 (1989)
27. DP McGinn, DH Johnson, Estimation of all-pole model parameters from noise-corrupted sequence. IEEE Transactions on Acoustics, Speech, and Signal Processing **37**(3), 433–436 (1989)
28. J Meyer, L Dentel, F Meunier, Speech recognition in natural background noise. PLoS One Journal **8**(11), 1–14 (2013)
29. G Muhammad, Noise-robust pitch detection using auto-correlation function with enhancement. Journal of King Said University- Computer and Information Sciences **22**, 13–28 (2010)
30. K Onoe, H Segi, T Kobayakawa, S Sato, T Imai, A Ando, *Filter bank subtraction for robust speech recognition*. Proceeding of ICSLP, 2002
31. H Rahali, Z Hajaiej, N Ellouze, Autocorrelation-domain method for noise robust speech recognition. International Journal of Tomography & Simulation **29**, 1 (2016)
32. S Seyedin, SM Ahadi, S Gazor, New features using robust MVDR spectrum of filtered autocorrelation sequence for robust speech recognition. The Scientific World Journal **2013**, Article ID 634160 (2013)
33. BJ Shannon, KK Paliwal, Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition. Speech Communication **48**, 1458–1485 (2006)
34. SPIB, *SPIB Noise Data*, 1995. Available from http://spib.linse.ufsc.br/noise.html
35. SV Vaseghi, Advanced digital signal processing and noise reduction, *4rd edn*. (Wiley Ltd., UK, 2008)
36. K-H Yuo, H-C Wang, Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences. Speech Communication **28**, 13–24 (1999)