Noise Estimation Using Mean Square Cross Prediction Error for Speech Enhancement

Gang Wang, Chunguang Li, and Le Dong, *Member, IEEE*

***Abstract*—This paper shows the feasibility of noise extraction from noisy speech and presents a two-stage approach for speech enhancement. The preproposed mean square cross prediction error (MSCPE) based blind source extraction algorithm is utilized to extract the additive noise from the noisy speech signal in the first stage. After that a modified spectral subtraction and a modified Wiener filter approach are proposed to extract the speech signal for speech enhancement in the second stage, where all the frequency spectra of the extracted noise are utilized. Theoretical justification shows that the MSCPE-based algorithm can extract desired signal from mixed sources. Experimental results show that the averaged correlation coefficient between the extracted noise and the original additive noise are beyond 85% for Gaussian noise and beyond 75% for real-world noise at $\mathrm{SNR} = 0\,\mathrm{dB}$, and the proposed speech enhancement approaches perform better than conventional methods, such as spectral subtraction and Wiener filter.**

***Index Terms*—Autoregressive (AR) parameter, blind source extraction (BSE), mean square cross prediction error (MSCPE), spectral subtraction (SS), speech enhancement, Wiener filter (WF).**

I. Introduction

SPEECH enhancement may be applied to mobile communication systems, speech recognition systems, or hearing aid devices, to mention just a few. The main objective of speech enhancement is to improve the performance of speech communication system in noisy environment. Over the past four decades, much research has focused on this area [1]–[17].

There are several types of speech enhancement problems according to the type of noise source, the way the noise interacts with the clean speech signal, the number of microphones, and the nature of the communication systems. This paper deals with the enhancement of a speech signal corrupted by an additive noise in a single microphone. The additive noise may be modelled as white Gaussian or come from the real world. There have been many algorithms proposed for solving this problem. Most of the algorithms are implemented in the spectral domain using discrete Fourier transform (DFT), such as spectral subtraction (SS) [1]–[3] and Wiener filter (WF) approach [4], [5]. The WF approach assumes that both the clean speech and additive noise have a Gaussian distribution. Recent research [6], [7] showed that the probability density function (pdf) of the DFT coefficients of short time stationary clean speech signal might be better modeled by a super Gaussian distribution, and a corresponding minimum mean-square error (MMSE) based algorithm was presented for speech enhancement. Besides spectral domain methods, linear transform domain-based maximum of a posteriori (MAP) [8]–[11] approaches have been proposed [11]–[15], such as Karhumen–Loève transform (KLT), discrete cosine transform (DCT) [14], [15], and independent component analysis (ICA) [11].

The aforementioned algorithms are mostly focused on how to *directly* estimate speech signal. Since clean speech signal is not always active but the additive noise is, the estimation of clean speech becomes difficult when the signal-to-noise ratio (SNR) is low, e.g., below zero dB. Then most of the above algorithms work poorly with low SNRs. Note that the noise becomes dominant when the speech is inactive, thus the estimation of the additive noise is more straightforward. This motivates us to present an *indirect* method for speech enhancement. In this method, the noise is extracted in the first stage and clean speech is to estimate in the second stage.

If the noise is efficiently extracted from the noisy speech signal, the spectrum of the clean speech can be estimated by Wiener filtering or spectral subtraction. The more correlated the extracted noise and the original noise is, the better performance can be achieved. The decorrelation of the noise and the clean speech in the noisy speech is usually carried out in a signal subspace, by using approaches like KLT, DCT, and ICA. However, this kind of transformation cannot estimate the additive noise well. A blind source extraction (BSE) [18]–[24] based algorithm, developed from the second order ICA algorithm (called AMUSE, or TICA algorithm [20]), is used here to extract the noise.

In this work we aim to show the feasibility of the noise extraction problem, and propose a two-stage approach for speech enhancement. Firstly, the noise is extracted using the mean square cross prediction error (MSCPE) algorithm. Secondly, the clean speech is estimated by using information of the extracted noise. In the first stage, the key assumption of extraction is that the AR parameter of the additive noise signal can be estimated. Since a voice activity detector (VAD) [17] can detect whether the speech is active or not, and the noise is dominant when the speech is not

Manuscript received July 31, 2009; revised March 26, 2010 and May 23, 2010; accepted May 25, 2010. Date of current version July 16, 2010. This work was supported by National Science Foundation of China under Grants 30900318 and 60871094. This paper was recommended by Associate Editor A. Uncini.

G. Wang is with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: wanggang_hld@hotmail.com).

C. Li is with the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: cgli@zju.edu.cn).

L. Dong is with the Institute of Intelligent Systems and Information Technology, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: ledongisi@gmail.com).

Digital Object Identifier 10.1109/TCSI.2010.2054930

active, the AR parameters of the noise can be estimated when the speech is not active. In the second stage, the spectra of the estimated noise are incorporated into Wiener filtering and spectral subtraction for speech enhancement.

It was demonstrated in [23] that a desired signal could be extracted from linear mixtures if its autoregressive (AR) parameters or linear predictor is known. Based on this finding, mean square prediction error (MSPE) based algorithms were proposed in noise-free case [23] and noisy case [24]. The MSPE-based algorithms are not robust enough to extract any desired signal using the corresponding AR parameters. Based on this observation, a new cost function based on MSCPE was proposed in [18], which caters for the specific AR model parameters and can extract any desired signal. Here an MSCPE-based algorithm is employed to extract the additive noise for speech enhancement.

This paper is organized as follows. In Section II, we present the framework of how to utilize BSE-based noise estimation algorithm to extract additive noise. In Section III, we introduce the preproposed MSCPE algorithm, and show the feasibility of the additive noise extraction problem in noisy speech. In Section IV, we propose a modified spectral subtraction and a modified Wiener filtering algorithm for speech enhancement. The numerical results for the speech data are described and analyzed in Section V. Discussions and conclusions are given in Section VI.

II. BSE-Based Noise Estimation Framework in Speech Enhancement

In this section, we establish a noise estimation model for speech enhancement. Let $s(t)$ be the clean speech signal. A K-dimension vector of samples of $s(t)$ at time n is denoted by $\boldsymbol{s}(n)$. Let $\boldsymbol{x}(n)$ and $\boldsymbol{v}(n)$ be the corresponding K-dimension vector of noisy speech signal and additive noise, respectively. We denote

$$\begin{aligned}\boldsymbol{v}(n) &= [v(n), v(n-1), \cdots, v(n-\mathrm{K}+1)]^{\boldsymbol{T}}\\ \boldsymbol{s}(n) &= [\mathrm{s}(n), \mathrm{s}(n-1), \cdots, \mathrm{s}(n-\mathrm{K}+1)]^{\boldsymbol{T}}\\ \boldsymbol{x}(n) &= [x(n), x(n-1), \cdots, x(n-\mathrm{K}+1)]^{\boldsymbol{T}} \qquad (1)\end{aligned}$$

where $(.)^T$ represents the transpose operation. When $\boldsymbol{v}(n)$ plays the role of desired signal and $\boldsymbol{s}(n)$ play the role of noise, the expression (1) can be regarded as a noisy BSE model

$$\boldsymbol{x}(n) = \boldsymbol{v}(n) + \boldsymbol{s}(n). \qquad (2)$$

The goal of desired signal estimation in BSE is to find a vector $\boldsymbol{w}$ such that $y(n) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}(n)$ is an estimated signal up to a scalar, which is estimated by SNR. In (1), it is not assumed that the $\boldsymbol{s}(n)$ and $\boldsymbol{v}(n)$ are normalized to be unit variance, since their variances are determined by SNR. In the BSE problem, the extracted signal usually has zero means and unit variance; therefore, the SNR in noisy speech has to be estimated. Here we utilized a VAD-based method [17] to estimate the SNR.

Remark 1 (Why Use BSE Model): Generally $\boldsymbol{v}$ can be regarded as white Gaussian noises which satisfy the following relations: $E\{v(n)v(n-\tau)\} = 0, 0 < \tau$. Thus, $\boldsymbol{v}$ with different delay can be regarded as different signals. In addition, the clean speech $\boldsymbol{s}$ plays the role of noise; therefore, $\boldsymbol{x}(n) = \boldsymbol{v}(n) + \boldsymbol{s}(n)$ can be regarded as a noisy BSE model. The general noisy BSE form is $\boldsymbol{x}(n) = \boldsymbol{A}\boldsymbol{v}(n) + \boldsymbol{s}(n)$, where $\boldsymbol{A}$ is an unknown mixing matrix. Here $\boldsymbol{A}$ is a unit matrix.

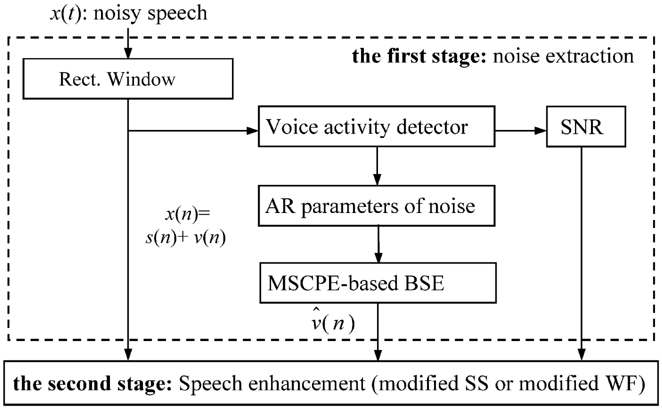


Fig. 1. Block diagram of the proposed system.

Remark 2 (Why Not Use Noisy BSE Algorithm): In general, there are two methods [19] to deal with noisy BSE problem. One is to neglect the noise in order to obtain tractable and simple results. The other is noise removal technique, in which the noise-free BSE algorithms are modified to remove, or at least reduce, noise. The second method seems promising, but it is based on the assumption that the additive noise has Gaussian distribution. Since the additive noise in our model is speech signal, which does not have Gaussian but super-Gaussian distribution, the second method cannot work well. Therefore, the noise-free BSE algorithm is adopted to extract desired signal $v(n)$.

The estimated noise signal should be correlated to the original noise and uncorrelated to the speech signal in order to obtain good performance of speech enhancement.When the speech is inactive, the BSE model (1) is degraded to $\boldsymbol{x}(n) = \boldsymbol{v}(n)$. If these additive noises, $v(n), v(n-1), \cdots, v(n-\mathrm{K}+1)$, such as Gaussian noises, are uncorrelated from each other, we can expect that the MSCPE-based algorithm will extract a noise $v(n+1-i)(0 < i < K+1)$ up to a scalar. This then provides full spectrum of desired noise. In addition, if these additive noises, such as airport noises, are correlated to some extent, the extracted signal may be a linear combination of these source signals. The extracted frequency spectra are inevitably distorted. When the speech is active, the BSE model would be $\boldsymbol{x}(n) = \boldsymbol{v}(n) + \boldsymbol{s}(n)$. Then the extracted signal would be polluted by speech data. The corresponding frequency spectra would be distorted to some extent. Thus, it is necessary to test the feasibility of noise extraction.

Remark 3 (How to Test the Feasibility of Noise Extraction): The following measure is used to test the feasibility of noise extraction from noisy speech: If the extracted noise helps to improve the performance of speech enhancement, the extraction algorithm can be regarded as *attractive*, otherwise *average*.

Altogether, the proposed BSE-based noise estimation framework is summarized in Fig. 1. The noisy speech signal is divided into frames with the length of 256 by rectangular windows, and the consecutive frames have an overlap of 128 points. The one dimension noisy speech data $x(n)$ are expanded to K-dimension vector $\boldsymbol{x}(n)$ as expression (1). Then the SNR and the AR parameters of the addtive noise are estimated using VAD, and the noise is estimated in the first stage. Secondly, the clean speech is estimated by modified SS or modified WF approach. It should be mentioned here that when the SNR of the noisy speech and the additive noise is stationary, the calculation of AR parameters and SNR is not needed in each frame. In practice, we calculate them in every one hundred frames.

III. MSCPE-Based BSE Algorithm

In this section, we show the principles of MSCPE-based noise estimation algorithm proposed in [18]. This algorithm is a noise-free BSE method, which will serve as the noise estimator for the first stage. Here we introduce the MSCPE-based BSE approach and give the comparison with the MSPE algorithm.

A. Cost Function of AR Parameter-Based BSE Algorithm

In BSE problem, we observe a K-dimensional stochastic signal vector $\boldsymbol{x}(n)$ of the form of $\boldsymbol{x}(n) = \boldsymbol{A}\boldsymbol{s}(n)$, where $\boldsymbol{A}$ is the mixing matrix and $\boldsymbol{s}(n)$ is the source signal vector. The goal of BSE is to find a vector $\boldsymbol{w}$ such that $y(n) = \boldsymbol{w^T}\boldsymbol{x}(n)$ is an estimated source signal up to a scalar. To cope with ill-conditioned case and to make algorithms simpler and faster, prewhitening is often used to transform the observed signals $\boldsymbol{x}$ to $\boldsymbol{Hx}$, such that $\boldsymbol{H}E\{\boldsymbol{x}\boldsymbol{x}^T\}\boldsymbol{H}^T = \boldsymbol{I}$, where $\boldsymbol{H}$ is a prewhitening matrix. For convenience, we assume that $\boldsymbol{x}$ has been prewhitened and has the same dimensions as $\boldsymbol{s}$.

Suppose that the AR parameters of the noise are known and denote the length of the AR model by p. The instantaneous prediction error (PE) denoted by $e(n)$ is as follows:

$$\begin{aligned} e(n) &= y(n) - \boldsymbol{b}^T\boldsymbol{Y}(n)\\ \boldsymbol{b} &= [b_1, b_2, \cdots, b_{\mathrm{P}}]^T\\ \boldsymbol{Y}(n) &= [y(n-1), y(n-2), \cdots, y(n-p)]^T\\ y(n) &= \boldsymbol{w}^T\boldsymbol{x}(n) \end{aligned} \tag{3}$$

where $\boldsymbol{b}$ is the AR parameter of a desired signal.

The mean cross prediction error (MCPE) [18] of output y can be expressed as $E\{e(n)e(n-q)\}$, where q denotes error delay. Then the corresponding $e(n)$ is as follows:

$$\begin{aligned} e(n) &= y(n) - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}(n)\\ &= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}(n) - [b_1, b_2, \cdots, b_P]\\ &\quad \times \left[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}(n-1), \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}(n-2), \cdots, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}(n-p)\right]^{\mathrm{T}}\\ &= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}(n) - \boldsymbol{w}^{\mathrm{T}}\sum_{\mathrm{i}=1}^{\mathrm{P}} b_{\mathrm{i}}\boldsymbol{x}(n-i)\\ &= \boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{x}(n) - \sum_{\mathrm{i}=1}^{\mathrm{P}} b_i\boldsymbol{x}(n-i)\right). \end{aligned} \tag{4}$$

Assuming that

$$\boldsymbol{z}(n) = \left(\boldsymbol{s}(n) - \sum_{i=1}^{p} b_{\boldsymbol{i}}\boldsymbol{s}(n-i)\right) \tag{5}$$

we have

$$\begin{aligned} e(n) &= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}(n) = \boldsymbol{z}^{\mathrm{T}}(n)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}\\ e(n-q) &= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}(n-q) = \boldsymbol{z}^{\mathrm{T}}(n-q)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}. \end{aligned} \tag{6}$$

Thus, the MCPE, expressed as $E\{e(n)e(n-q)\}$, is

$$\begin{aligned} E\{e(n)e(n-q)\} &= E\left\{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}(n)\boldsymbol{z}^{\mathrm{T}}(n-q)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}\right\}\\ &= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}E\left\{\boldsymbol{z}(n)\boldsymbol{z}^{\mathrm{T}}(n-q)\right\}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}. \end{aligned} \tag{7}$$

Furthermore, assuming that

$$\boldsymbol{Z}(q) = E\left\{\boldsymbol{z}(n)\boldsymbol{z}^{\mathrm{T}}(n-q)\right\} \tag{8}$$

the MCPE can be expressed as

$$E\{e(n)e(n-q)\} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Z}(q)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}. \tag{9}$$

In [18], we proposed the mean square cross prediction error (MSCPE), expressed as $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Z}(q)\boldsymbol{Z}^{\mathrm{T}}(q)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}^{\mathrm{T}}$, as a cost function [18] to extract the desired signal. The cost function in a simply form is as follows:

$$\begin{aligned} J_q(\boldsymbol{w}) &= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Z}(q)\boldsymbol{Z}^{\mathrm{T}}(q)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}, \quad 0 < q \le p\\ \text{s.t.}\quad \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} &= 1. \end{aligned} \tag{10}$$

Denote $\boldsymbol{Z}(q)\boldsymbol{Z}^{\mathrm{T}}(q)$ by $\boldsymbol{\Sigma}$, then $\boldsymbol{\Sigma}$ is a diagonal matrix, whose diagonal element $\Sigma(i,i)(i = 1, 2, \ldots, M)$ is equal to the square of the MCPE of source signal $s_i(i = 1, 2, \ldots, M)$. If the source signals have different AR model parameters, MSCPE would have the unique minimum, i.e., zero. Thus, we can extract any desired signal by minimizing $J_q(\boldsymbol{w})$.

B. Algorithm Using Eigenvalue Decomposition (EVD) or Singular Value Decomposition (SVD)

Note that the expression (10) implies that minimizing the cost function $J_q(\boldsymbol{w})$ under the constraint $\boldsymbol{w^T}\boldsymbol{w} = 1$ is equivalent to finding the eigenvector corresponding to the minimal eigenvalue of the real symmetric matrix $\boldsymbol{A}\boldsymbol{Z}(q)\boldsymbol{Z}^{\mathrm{T}}(q)\boldsymbol{A}^{\mathrm{T}}$. Moreover, we could also find that $\boldsymbol{w}$ is equivalent to the singular vector of the minimal singular value of $\boldsymbol{Z}(q)$. Thus, we have the following algorithm [18]:

$$\begin{aligned} \boldsymbol{z}(n) &= \left(\boldsymbol{s}(n) - \sum_{i=1}^{p} b_{\boldsymbol{i}}\boldsymbol{s}(n-i)\right)\\ \boldsymbol{Z}(q) &= E\left\{\boldsymbol{z}(n)\boldsymbol{z}^{\mathrm{T}}(n-q)\right\}\\ \boldsymbol{w} &= \mathrm{MINEVD}\left\{\boldsymbol{A}\boldsymbol{Z}(q)\boldsymbol{Z}^{\mathrm{T}}(q)\boldsymbol{A}^{\mathrm{T}}\right\}\\ &= \mathrm{MINSVD}\left\{\boldsymbol{A}\boldsymbol{Z}(q)\boldsymbol{A}^{\mathrm{T}}\right\} \end{aligned} \tag{11}$$

where $\mathrm{MINEVD}\{\boldsymbol{M}\}$ is an operator that calculates the normalized eigenvector corresponding to the minimal eigenvalue of a real symmetric matrix $\boldsymbol{M}$, and $\mathrm{MINSVD}\{\boldsymbol{M}\}$ is an operator that calculates the normalized singular vector corresponding to the minimal singular value of $\boldsymbol{M}$.

C. Theoretical Justification of the Desired Property

The following theorem verifies the desired property of the MSCPE-based BSE algorithm.

Theorem 1: In the noise-free BSE problem, the observation $\boldsymbol{x}(n)$, has the form of $\boldsymbol{x}(n) = \boldsymbol{As}(n)$, where $\boldsymbol{A}$ is the unknown mixing matrix and $\boldsymbol{s}(n)$ is the source signals vector. Suppose that:

1) sources are not correlated with each other and have different temporal structures satisfying the following relations:

$$E\left\{s_i(n)s_j(n-\tau)\right\} = 0, \quad \forall i \neq j, 0 \leq \tau$$

2) $\boldsymbol{x}$ has been prewhitened;
3) different source signals have different AR parameters;
4) $\boldsymbol{x}$ has the same dimensions (for example, K) as $\boldsymbol{s}$.

If the AR parameters of any desired signal are known, the proposed algorithm (11) can extract the desired signal.

Proof: When condition 1) is satisfied, $\boldsymbol{Z}(q)$ becomes a diagonal matrix, and $\boldsymbol{Z}(q)\boldsymbol{Z}^{\mathrm{T}}(q)$ would be also a diagonal matrix whose diagonal elements are nonnegative. When $\boldsymbol{x}$ is prewhitened, i.e., $E\{\boldsymbol{xx}^T\} = \boldsymbol{I}$, $\boldsymbol{A}$ becomes an orthogonal matrix. Denoting $\Lambda = \boldsymbol{Z}(q)\boldsymbol{Z}^{\mathrm{T}}(q)$, the SVD of $\boldsymbol{AZ}(q)\boldsymbol{Z}^{\mathrm{T}}(q)\boldsymbol{A}^{\mathrm{T}}$ can be expressed by $[\boldsymbol{A}, \Lambda, \boldsymbol{A}^{\mathrm{T}}] = \mathrm{SVD}(\boldsymbol{AZ}(q)\boldsymbol{Z}^{\mathrm{T}}(q)\boldsymbol{A}^{\mathrm{T}})$.

For given AR model parameters $\boldsymbol{b}$ of the desired source signal s_k, where k is the desired number, the MCPE of each source is expressed as $E\{e_j(n)e_j(n-q)\}(j = 1, 2, \ldots, \mathrm{K})$ [18], where $e_j(n)$ is the residual error. Since source signals have different AR parameters, we have the following properties:

$$E\left\{e_j(n)e_j(n-q)\right\} = \begin{cases} = 0, & j = k \\ \neq 0, & j \neq k \end{cases}, \quad 0 < q \leq p$$

where

$$e_j(n-q) = s_j(n-q) - \sum_{i=1}^{p} b_i s_j(n-i-q), \quad q \geq 0$$

and p is the length of the AR parameters. Thus, the k-th diagonal element of Λ will be zero, which is the minimum value of the diagonal elements.

Minimizing the cost function $J_q(\boldsymbol{w})$ under the constraint $\boldsymbol{w^T w} = 1$ is equivalent to finding the eigenvector corresponding to the minimal eigenvalue of the real symmetric matrix $\boldsymbol{AZ}(q)\boldsymbol{Z}^{\mathrm{T}}(q)\boldsymbol{A}$. Thus, the proposed algorithm (11) can extract the desired signal. (End of Proof)

Remark 4 (Relationship Between MSPE and MSCPE-Based BSE Algorithm): It has been shown in [23] and [24] that source signals can be extracted successfully by minimizing the normalized MSPE $E\{e^2(n)\}/E\{y^2(n)\}$ as long as they have different temporal structures. As mentioned in Section III-A, it is assumed that $\boldsymbol{x}$ is prewhitened and thus the output power of the demixing scalar $E\{y^2(n)\}$ is unitary. Therefore, the cost function can be set as MSPE, i.e., $E\{e^2(n)\}$, which is written as follows:

$$E\left\{e^2(n)\right\} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{AR_p A^T w} \tag{12}$$

and

$$\begin{aligned} \boldsymbol{R_p} &= E\left\{\left(\boldsymbol{s}(n) - \sum_{\boldsymbol{k}=1}^{p} b_{\boldsymbol{k}}\boldsymbol{s}(n-i)\right) \times \left(\boldsymbol{s}(n) - \sum_{\boldsymbol{i}=1}^{p} b_{\boldsymbol{i}}\boldsymbol{s}(n-i)\right)^T\right\} \\ &= E\left\{\boldsymbol{z}(n)\boldsymbol{z}^{\mathrm{T}}(n)\right\} = \boldsymbol{Z}(q=0) \\ \boldsymbol{w} &= \mathrm{MINEVD}\left\{\boldsymbol{AZ}(q=0)\boldsymbol{A}^{\mathrm{T}}\right\} \\ &= \mathrm{MINSVD}\left\{\boldsymbol{AZ}(q=0)\boldsymbol{A}^{\mathrm{T}}\right\} \end{aligned} \tag{13}$$

where $\boldsymbol{R}_p$ is a diagonal matrix and its diagonal element $\boldsymbol{R}_p(i,i)$ equals to the MSPE $E\{\boldsymbol{e}_i^2(\mathrm{n})\}$ of the corresponding source signal. Moreover, when q is equal to zero, the MSCPE algorithm will degenerate to the MSPE algorithm.

D. Experiments of MCPE for Noise and Speech Signals

Comparing (10) with (13), we find that the MSPE-based algorithm extracts the source signal with the minimum MSPE as the first output, and the MSCPE-based algorithm extracts the source signal with the minimum square of MCPE. Simulations in [18] show that the MSCPE-based algorithm is more robust than the MSPE-based algorithm. Then we can estimate the desired noise signal through MSCPE algorithm if the square of MCPE of the additive noise is smaller than that of the clean speech. In the following experiment, we will calculate the MCPE of different signals given a desired noise signal's AR parameters in order to test the validity of this method.

The clean speech data, a male signal called "sp01" and a female signal called "sp30," are obtained from a noisy speech corpus (NOIZEUS) [25], [26], and noise signals are estimated by subtracting the clean speech from the corresponding noisy speech in NOIZEUS. Here we choose four types of noises for experiments, which are Gaussian, "airport," "babble," and "train." Gaussian noise is generated by the MATLAB function "randn." The noises are divided into a training set and a test set.

In this experiment, firstly we choose a desired noise which is associated with a background noise. Secondly, we estimate the AR parameters of the desired noise in the training set. And finally, we calculate the corresponding MCPE values of the corresponding noise in the test set and two clean speech signals using expression (7). The AR model parameters of desired noise are obtained using Matlab function "aryule" with a length of $p = 20$, and $q = 1$ in expression (11).

Table I shows that the absolute value of MCPE of noise, highlighted in italic and bold, is smaller than that of the clean speech. This means that the MSCPE algorithm extracts the additive noise as the first output if this kind of noise and clean speech are linearly mixed.

After the AR model parameter is estimated, the MSCPE algorithm (11) can be used to extract the noise, denoted by $\hat{v}(n)$, from noisy speech signal, and we have $\hat{v}(n) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{Hx}(n)$, where $\boldsymbol{H}$ is a prewhitening matrix.

TABLE I
MCPE VALUES OF DESIRED NOISE AND TWO SPEECH SIGNALS WHEN THE AR PARAMETERS ARE CALCULATED THROUGH THE CORRESPONDING NOISE IN THE TRAINING SETS

MCPE	Speech 1(male)	Speech 2(female)	Noise
Gaussian	0.8065	0.6988	***-0.0060***
Airport	0.0673	0.0073	***0.0048***
Babble	0.1094	0.0588	***0.0389***
Train	0.3029	0.2684	***0.0100***

IV. MSCPE-Based Spectral Subtraction and Wiener Filter

When the estimated noise $\hat{v}(n)$ is extracted, we can utilize some conventional algorithms to enhance speech. We here introduced two modified approaches using spectral subtraction and the Wiener filter for the second stage.

A. Modified Spectral Subtraction (MSS)

The main idea of spectral subtraction is to estimate the magnitude frequency spectrum of the underlying clean speech by subtracting the noise magnitude spectrum from the noisy speech spectrum. Note that $x(n)$, $s(n)$, and $v(n)$ are a one-dimensional vector of noisy speech, clean speech, and additive noise, respectively. Let $X(\omega)$, $S(\omega)$, and $V(\omega)$ be the short-time spectra associated with the windowed time function $x(n)$, $s(n)$, and $v(n)$. Then we have

$$x(n) = s(n) + v(n). \tag{14}$$

The corresponding Fourier transformation is

$$X(\omega) = S(\omega) + V(\omega). \tag{15}$$

Let $\hat{V}(\omega)$ be the spectrum of estimated noise $\hat{v}(n)$. Then the spectral subtraction filter $H(\omega)$ is calculated by replacing the noise spectrum $V(\omega)$ with the spectra of estimated noise. The magnitude $|V(\omega)|$ of $V(\omega)$ is replaced by the magnitude $|V(\omega)|$ of $\hat{V}(\omega)$, and the phase $\theta_V(\omega)$ of $V(\omega)$ is replaced by the phase $\theta_X(\omega)$ of $X(\omega)$. These substitutions results in the modified spectral subtraction estimator $\hat{S}(\omega)$

$$\begin{aligned}\hat{S}(\omega) &= \left[|X(\omega)| - \left|\hat{V}(\omega)\right|\right] e^{j\theta}\\ \theta &= \theta_X(\omega)\end{aligned} \tag{16}$$

or

$$\begin{aligned}\hat{S}(\omega) &= H(\omega)X(\omega)\\ H(\omega) &= 1 - \frac{\left|\hat{V}(\omega)\right|}{|\mathrm{X}(\omega)|}.\end{aligned} \tag{17}$$

Moreover, $H(\omega)$ is set to zero if $|X(\omega)|$ is less than $|\hat{V}(\omega)|$. Finally, we obtain the enhanced speech signal $\hat{s}(n)$ by inverse DFT of $\hat{S}(\omega)$.

The difference between this modified spectral subtraction approach and those conventional methods [1]–[3] lies in that the magnitude $|V(\omega)|$ of $V(\omega)$ is not replaced by the average value taken during nonspeech activity, but by the magnitude $|\hat{V}(\omega)|$ of $\hat{V}(\omega)$.

B. Modified Wiener Filter (MWF)

The main idea of the Wiener filter is to constructs an "optimum" filter under the assumption that both the clean speech and noise have Gaussian distributions. Let $P_X(\omega)$, $P_S(\omega)$, $P_V(\omega)$ and $\hat{P}_V(\omega)$ be the short-time power density spectra associated with the windowed time function $x(n)$, $s(n)$, $v(n)$ and $\hat{v}(n)$. Since we have obtained the estimated noise $\hat{v}(n)$, the modified Wiener filter estimator $\hat{S}(\omega)$ is

$$\begin{aligned}\hat{S}(\omega) &= H(\omega)X(\omega)\\ H(\omega) &= \frac{P_S(\omega)}{P_S(\omega) + P_V(\omega)}\\ &\approx \frac{P_X(\omega) - \hat{P}_V(\omega)}{P_{X(\omega)}}\\ &= \frac{|X(\omega)|^2 - \left|\hat{V}(\omega)\right|^2}{|X(\omega)|^2}.\end{aligned} \tag{18}$$

Moreover, $H(\omega)$ is set to zero if $P_X(\omega)$ is less than $\hat{P}_V(\omega)$. Finally, we obtain the enhanced speech signal $\hat{s}(n)$ by inverse DFT of $\hat{S}(\omega)$.

The difference between the modified Wiener filter approach and classical ones [4], [5] lies in that the magnitude $|P_V(\omega)|$ of $P_V(\omega)$ is not replaced by the average value taken during nonspeech activity, but by the magnitude $|\hat{P}_V(\omega)|$ of $\hat{P}_V(\omega)$.

If the extracted noise is correlated with the original noise, the estimated spectra of noise will contain more information than the average value or the variance of the estimated noise. We can expect that the modified spectral subtraction and the modified Wiener filter approach will perform better than the conventional ones when the MSCPE algorithm extracts the desired noise.

V. Experimental Results

The noisy speech signal has the sampling frequency of 8000 Hz, and its length of samples approaches 21 000. In the first stage, the noisy speech signal is divided into frames with the length of 256, and the consecutive frames have an overlap of 128 points. Then one-dimensional noisy speech data $x(n)$ are expanded to K-dimensional vector $\boldsymbol{x}(n)$ as (1). The AR parameter $\boldsymbol{b}$ of a desired noise has the length of p. Here we adopted that $K = 5, p = 20$ and $q = 1$ in (10). From a theoretical point of view, $q > 1$ will not change the order of source estimation. However, in our test, large q will demand large K, the dimension of the BSE model, thus increasing the calculation of further SVD. We adopt $q = 1$ in this paper.

The estimated noise signal has the form of $\hat{v}(n) = \boldsymbol{w}^{\mathbf{T}}\boldsymbol{H}\boldsymbol{x}(n)$. Note that $\boldsymbol{H}$ is a prewhitening matrix, then $\hat{v}(n)$ has zero mean and unit variance. Thus, $\hat{v}(n)$ is multiplied by a scalar according to the SNR. When the inactivity of the voice is detected by VAD, the AR parameters and SNR are estimated in the inactive episode.

In the second stage, noisy data are transformed into frequency spectrum as the above frames using a rectangular window. Then modified spectral subtraction and modified

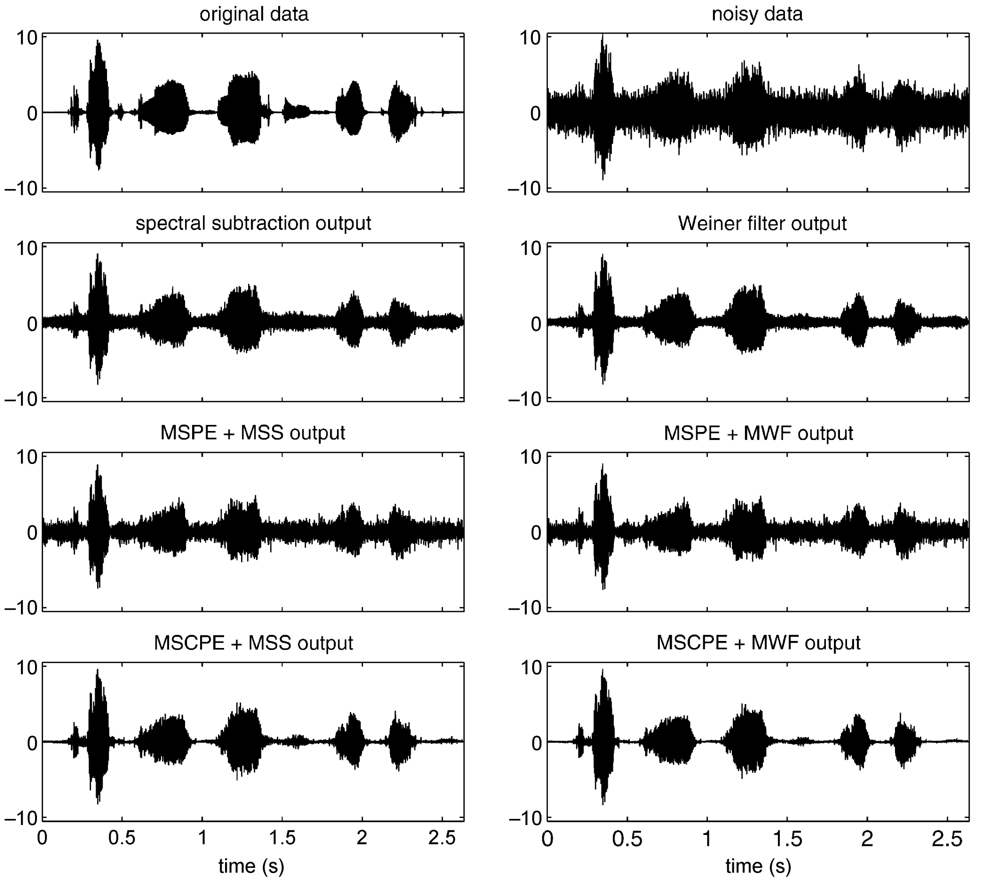


Fig. 2. Waveforms of male speech (“sp01”) corrupted by Gaussian noise at $\mathrm{SNR} = 0\ \mathrm{dB}$.

Wiener filter approaches are utilized to enhance the speech. The overlapping part should be averaged. The enhanced signal is denoted by $\hat{s}(n)$.

Regarding the performance index, we evaluate the correlation between the extracted signal and the original noise in the first stage. This performance is estimated based on the correlation coefficient (CC) between the extracted signal denoted by s_e and the original signal s_o; that is,

$$\mathrm{CC} = \left| \langle s_0, s_e \rangle / \sqrt{\langle s_0, s_0 \rangle \langle s_e, s_e \rangle} \right| \tag{19}$$

where operator $\langle\ \rangle$ denotes inner product, and CC is between 0 and 1.

We use spectral distortion (SD) and output SNR as criteria [6] for the performance of speech enhancement for the second stage. The SD between $x(n)$ and $y(n)$ with length N is calculated as follows. First, $x(n)$ and $y(n)$ are normalized to be zero mean and unit variance. Second, $x(n)$ and $y(n)$ are both divided into frames of length 64 samples without overlapping. Third, the 256-point fast Fourier transform (FFT) is calculated for each frame after padding 192 zeros into each frame. Let $X(k)$ and $Y(k)$ be the spectra associated with the windowed time function $x(n)$ and $y(n)$. Finally, the SD in decibels is defined as

$$\mathrm{SD}(x(n), y(n)) = \frac{1}{4N} \sum_{\boldsymbol{i}=1}^{N/64} \sum_{k=0}^{255} 20 \left| \log_{10} |X(k)| - \log_{10} |Y(k)| \right|. \tag{20}$$

The output SNR of the enhanced signal $\hat{s}(n)$ is defined as

$$\text{ouput SNR} = 10 \log_{10} \frac{\sum\limits_{n=1}^{N} s^2(n)}{\sum\limits_{n=1}^{N} \left(s(n) - \hat{s}(n)\right)^2} \tag{21}$$

where $s(n)$ is the clean speech signal.

It is easy to know that the higher CC or SNR is, the better the performance is. The lower SD is, the better the performance is.

A. *Performance in Gaussian Noise*

As mentioned in Section III, the clean speech data, a male signal called “sp01” and a female signal called “sp30,” are obtained from a noisy speech corpus (NOIZEUS) [25], [26], and the Gaussian noise is generated using MATLAB function “randn.” The noise is added to the original speech signal with different SNRs, namely −5, 0, 5, and 10 dB. These SNRs are assumed to be known. Denote spectral subtraction by SS, and the Wiener filter by WF. For comparison between MSPE and MSCPE, we enhanced the noisy speech using six methods, namely SS, WF, MSPE+MSS, MSPE+MWF, MSCPE+MSS, and MSCPE+MWF.

An example of clean, noisy, and enhanced male speech signals at 0 dB are depicted in Fig. 2. It is observed in the time domain that the enhanced speech signals obtained by modified ones utilizing MSCPE have a much lower residual noise level than other methods. However, the modified methods utilizing

TABLE II
COMPARISON OF OUTPUT SNR (IN dB) OF ENHANCED SIGNAL UNDER GAUSSIAN NOISE CORRUPTION

	Enhancement algorithm	Gaussian noise			
		-5 dB	0 dB	5 dB	10dB
Male	SS	1.03	4.14	8.37	12.3
	WF	1.04	4.21	8.48	12.5
	MSPE+MSS	0.79	3.57	7.56	11.9
	MSPE+MWF	0.85	3.78	7.78	12.1
	MSCPE+MSS	5.05	8.49	12.2	15.9
	MSCPE+MWF	5.62	8.88	12.5	16.0
Female	SS	1.00	4.05	7.96	12.1
	WF	1.01	4.17	7.97	12.3
	MSPE+MSS	0.75	3.51	7.12	11.5
	MSPE+MWF	0.81	3.67	7.21	11.6
	MSCPE+MSS	3.93	7.21	10. 7	14.5
	MSCPE+MWF	4.10	7.35	10.8	14.6

TABLE III
COMPARISON OF SPECTRAL DISTORTION OF NOISY SIGNALS (IN dB) OF ENHANCED SIGNAL UNDER GAUSSIAN NOISE CORRUPTION

	Enhancement algorithm	Gaussian noise			
		-5 dB	0 dB	5 dB	10dB
Male	Input SD	11.4	10.6	9.27	7.71
	SS	9.32	8.32	7.56	5.78
	WF	9.30	8.35	7.64	5.65
	MSPE+MSS	10.12	9.11	8.45	6.34
	MSPE+MWF	10.22	8.98	8.43	6.21
	MSCPE+MSS	6.84	5.51	4.59	3.95
	MSCPE+MWF	5.97	4.90	4.07	3.75
Female	Input SD	9.68	8.88	7.65	6.21
	SS	7.65	6.65	5.76	4.29
	WF	7.63	6.43	5.54	4.21
	MSPE+MSS	8.14	7.11	6.09	5.31
	MSPE+MWF	8.11	7.08	6.05	5.09
	MSCPE+MSS	5.58	4.85	4.01	3.41
	MSCPE+MWF	5.31	4.66	3.99	3.56

MSPE have slightly higher residual noise levels than the original SS and WF methods.

The obtained output SNRs and SDs are summarized in Table II and Table III, respectively. The results are averaged over 100 independent runs. It shows that two modified approaches utilizing MSCPE obtain better performance than other methods. The output SNRs for male and female speech are both improved. The output SDs are reduced to nearly the half of the input SDs in our algorithms. In addition, MSPE-based methods do not perform better than convention methods.

With respect to the comparisons among different speech signals, Tables II and III show that the modified algorithms utilizing MSCPE work slightly better in the male speech for a certain input SNR. This is for that CC between the estimated noise extracted from male noisy speech, and the original noise is higher than CC from female noisy speech. Fig. 3 plots the associated CCs at different SNR. The results are averaged over 100 independent runs. The length of estimated noise is not the frame length, but the same as the length of the signal in Fig. 2. It shows that CCs for male (circle) speech is slightly higher than CCs for female (triangle) speech. It also indicates that under the same input SNR the higher CC is, the better the performance is.

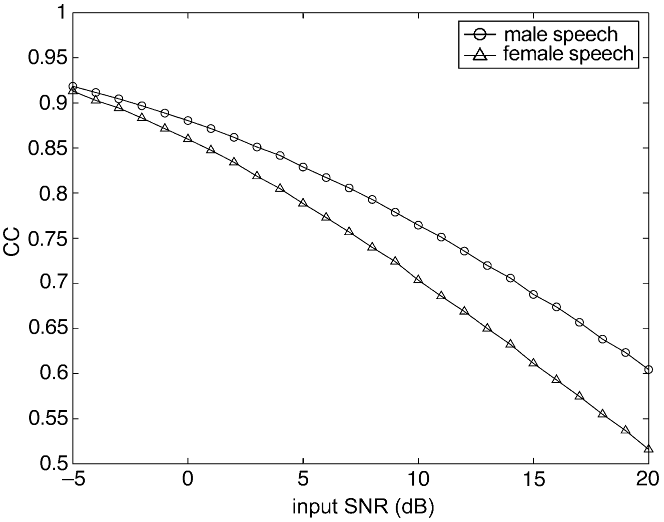


Fig. 3. CCs between the additive noises extracted by MSCPE algorithm and the original noises at different SNRs. The speech signals include female (triangle) and male (circle) ones.

When the estimated noise signal is divided into nonoverlapping consecutive frames with length 256, the corresponding CCs for female speech at input $\mathrm{SNR} = 0$ in one example is shown in Fig. 4. Besides CC values, the original male speech and the associated noisy speech are also given. Comparing the shape of the three waveforms, we find that CC obtained by MSCPE is approximate to one when the original speech is inactive. It means that the extracted noise by MSPCE is greatly correlated with the original noise, and the degraded BSE model $\boldsymbol{x}(n) = \boldsymbol{v}(n)$ does work. CC is relatively small when the original speech is active, which is corresponding to the BSE model $\boldsymbol{x}(n) = \boldsymbol{v}(n) + \boldsymbol{s}(n)$. Most of the CC values obtained by MSCPE are above 0.6. Fig. 3 shows that the average CC at $\mathrm{SNR} = 0$ is 0.88 for male and 0.87 for female speech. This means that the MSCPE algorithm works well in extracting the noise in speech enhancement. In addition, Fig. 4 also shows that CC obtained by MSPE is below 0.5. As an output result, Fig. 2 shows that MSPE+MSS and MSPE+MWF do not work as well as MSCPE+MSS and MSCPE+MWF. This means that the MSPE algorithm cannot help enhance speech. Therefore, the MSCPE extraction algorithm is *attractive* while MSPE is *average*.

B. Performance in Real World Noise

We still utilized a male signal "sp01" and a female signal "sp30" as the clean speech data. Noises are estimated by subtracting the clean speech from the corresponding noisy speech in NOIZEUS. Here we chose aforementioned "airport," "babble," and "train" noises for experiments. The noise is added to the original speech signal with different SNRs, namely −5, 0, 5, and 10 dB. Varied from Gaussian noise, the three noises are not stationary. Their variances are estimated by a minimum statistics method [16]. We enhanced the noisy speech using the aforementioned four methods in Section V. A, namely SS, WF, MSS, and MWF. Here $\mathrm{K} = 5$, $p = 20$ and $q = 1$ in (10). The noises are all extracted by MSCPE algorithm.

The results for real-world noise in terms of output SNR and SD are given in Tables IV and V, respectively. It is observed that

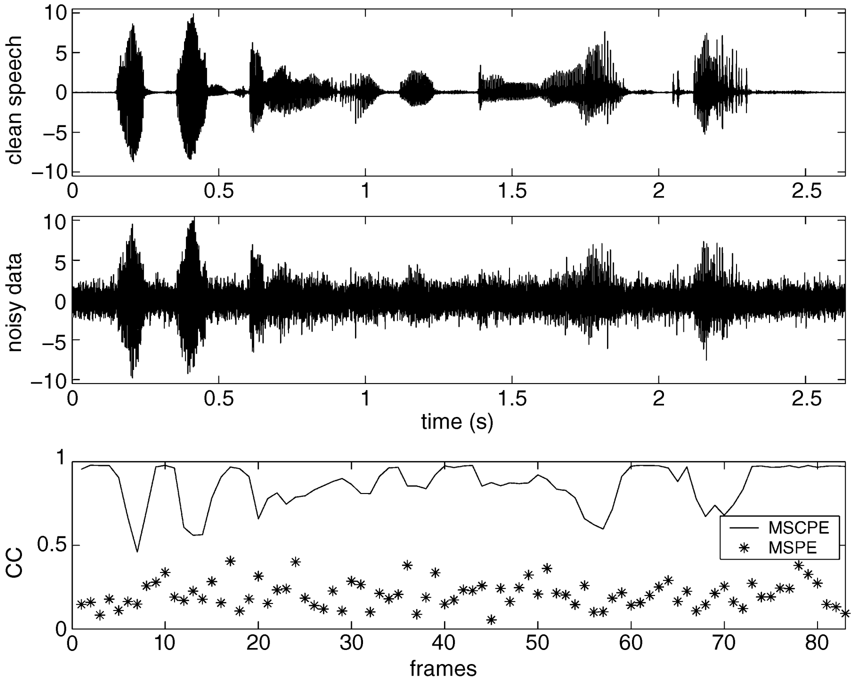


Fig. 4. CCs in consecutive frames between the signal extracted from noisy female speech and the original noise at input $\mathrm{SNR} = 0$.

TABLE IV
COMPARISON OF OUTPUT SNR (IN dB) OF ENHANCED SIGNAL UNDER AIRPORT, BABBLE, AND TRAIN NOISE CORRUPTION (THE NOISES ARE EXTRACTED BY THE MSCPE ALGORITHM)

	Algorithm	Airport noise (dB)				Babble noise (dB)				Train noise (dB)			
		-5	0	5	10	-5	0	5	10	-5	0	5	10
Male	SS	-0.10	3.1	6.5	11.3	-0.065	3.3	6.5	11.3	-0.053	2.8	6.8	11.1
	WF	-0.11	3.2	6.8	11.2	-0.054	3.2	6.7	11.2	-0.054	2.9	6.7	11.2
	MSS	0.79	4.7	8.6	12.7	0.74	4.7	8.7	12.8	0.39	4.8	8.8	13.1
	MWF	0.88	5.0	9.0	13.0	0.79	5.1	9.3	13.2	0.59	5.3	9.5	13.4
Female	SS	0.12	2.8	6.1	10.9	0.14	3.2	6.2	10.9	0.099	2.7	6.5	10.9
	WF	0.10	2.9	6.2	10.8	0.16	3.3	6.4	10.9	0.089	2.6	6.4	10.8
	MSS	1.20	4.5	8.0	12.1	1.1	4.4	8.2	12.3	0.63	4.2	8.3	12.5
	MWF	1.45	4.6	8.2	12.2	1.5	4.7	8.5	12.6	1.0	4.6	8.7	12.8

TABLE V
COMPARISON OF SPECTRAL DISTORTION OF NOISY SIGNALS (IN dB) OF THE ENHANCED SIGNAL UNDER AIRPORT, BABBLE, AND TRAIN NOISE CORRUPTION

	Algorithm	Airport noise (dB)				Babble noise (dB)				Train noise (dB)			
		-5	0	5	10	-5	0	5	10	-5	0	5	10
Male	Input SD	8.3	7.6	6.5	5.2	8.4	7.7	6.6	5.3	9.8	9.0	7.8	6.4
	SS	8.3	7.6	6.5	5.2	8.4	7.6	6.5	5.3	9.8	9.0	7.8	6.4
	WF	8.3	7.5	6.4	5.2	8.4	7.6	6.5	5.2	9.8	9.0	7.8	6.4
	MSS	6.5	5.3	4.3	3.6	7.0	5.7	4.8	4.0	8.1	6.7	5.8	4.5
	MWF	6.3	5.1	4.2	3.5	7.0	5.5	4.6	3.9	8.1	6.5	5.7	4.4
Female	Input SD	6.8	6.1	5.1	4.0	6.9	6.2	5.1	4.1	8.1	7.4	6.2	4.9
	SS	6.6	6.0	4.8	3.7	6.8	6.0	5.0	4.0	8.0	7.2	6.1	4.8
	WF	6.5	6.0	4.7	3.7	6.8	6.1	5.0	4.0	8.0	7.2	6.0	4.7
	MSS	5.6	4.7	3.9	3.1	5.8	4.9	4.1	3.5	7.3	5.9	4.9	4.1
	MWF	5.7	4.9	4.2	3.3	5.9	4.9	4.2	3.6	7.4	5.7	5.1	4.4

the proposed MSS and MWF algorithms perform better than spectral subtraction and Wiener filter for both male and female speakers at both output SNRs and SDs for the real-world noise.

Tables IV and V show that the modified algorithms work slightly better in the female speech at $\mathrm{SNR} = -5$, while they work slightly better in the male speech at other SNRs (0, 5, and 10). This is also caused by that the more correlation between the estimated noise and the original noise is, the better performance is achieved. Fig. 5 plots the associated CCs at different SNR. It shows that CCs for female (point) speech is slightly higher than CCs for male (solid line) speech when the SNR is below five. Both Fig. 3 and Fig. 5 show that the lower the SNR is, the higher CC is.

When the estimated noises are divided into consecutive frames (without overlap) with length 256, the corresponding CCs for female speech in one example is shown in Fig. 6. We

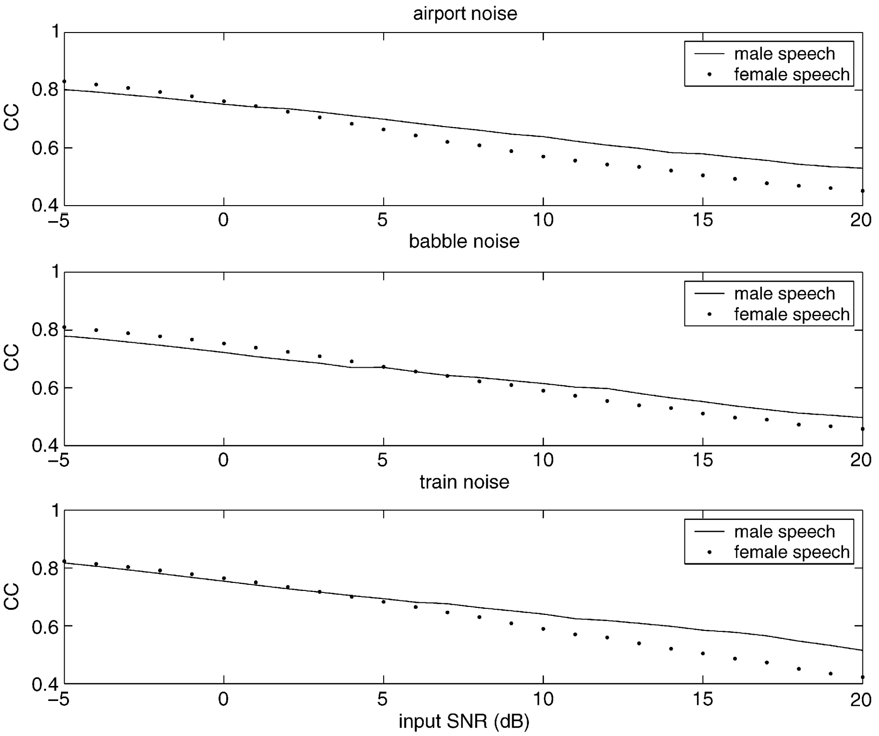


Fig. 5. CCs between the signal extracted from noisy speech and the original real-world noise at different SNRs. The speech signals include female (triangle) and male (circle) ones.

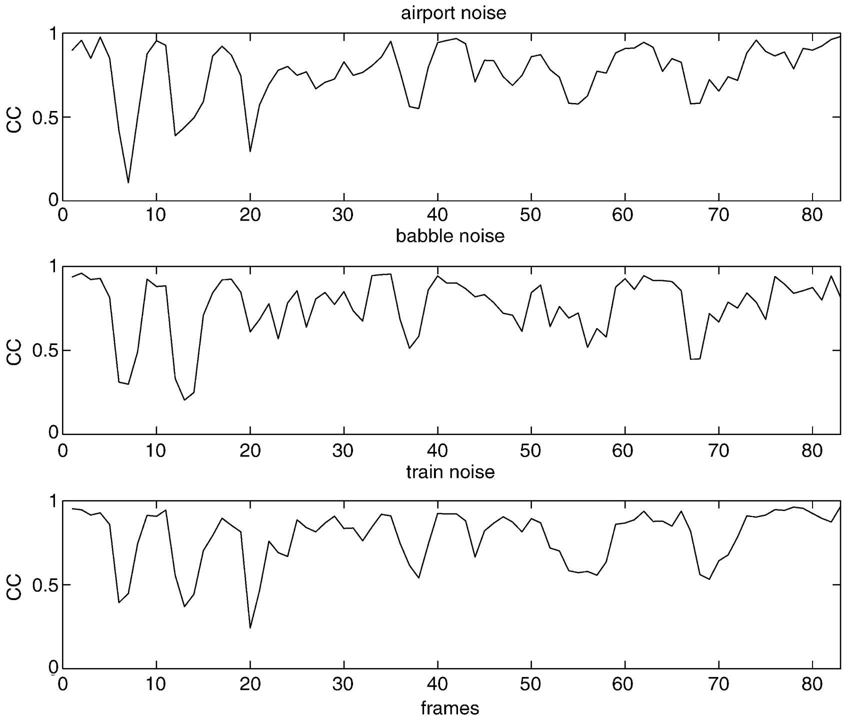


Fig. 6. CCs in consecutive frames between the extracted noises from noisy female speech ("sp30") and the original (airport, babble, and train) noise at input $\mathrm{SNR} = 0$.

also find that CC approaches one when the original speech is inactive. This means the extracted noise is strongly correlated with the original noise, and we can obtain most of the frequency spectra. Then the modified algorithms can enhance the speech. On the other hand, CC is relatively small when the original speech is active. Most of the CC values are above 0.5. Altogether, the modified algorithms also enhance the speech, which can be measured by output SNR and SD. For the airport noise at $\mathrm{SNR} = 0$, Fig. 5 shows that the averaged CC is 0.78. The output SNR shown in Table IV is increased by about 5 dB, and the SD shown in Table V is deceased from 7 dB to 5 dB.

For the real-world noise, the improvement of the output SNR is lower than that for the Gaussian noise. It can be explained as follows. The real-world noises with different time delays are usually correlated with each other. This would distort the frequency spectra of the noise and decrease the performance of speech enhancement. From the point of view of improving the performance of speech enhancement, the MSCPE extraction algorithm is also *attractive*.

In the experiments, we also test other clean speech data by using "sp02," "sp03," ... "sp29" obtained from NOIZEUS. Similar results are obtained. Owing to the limited length of the paper, these results are not shown here.

VI. Discussion and Conclusions

In this paper, we have shown the feasibility of the noise extraction problem in noisy speech and proposed a two-stage approach to solve the speech enhancement problem. In the first stage, the noise is extracted from the expanded noisy speech signals using an MSCPE-based BSE algorithm. In the second stage, a modified spectral subtraction and a modified Wiener filter approach are proposed to extract the speech signal. The main difference between the proposed method and the conventional methods [1]–[5] is that we do not utilize the average value or the variance of the estimated noise, but all spectra parameters of the estimated noises.

Both theoretical justification and experimental results show that MSCPE algorithm works better than MSPE in extracting desired signal. When MSCPE algorithm is utilized to extract additive noise, experiments show that the averaged CC between the extracted noise and the original additive noise are beyond 85% for Gaussian noise and beyond 75% for real-world noise at $\mathrm{SNR} = 0$ dB, and the lower the SNR is, the higher CC is. When MSPE algorithm is utilized, the averaged CC is below 50%.

The implementation of the estimators based on MSCPE requires the computation of the EVD or SVD twice in the first stage. One is for prewhitening, and the other is for optimization of MSCPE. In the second stage, FFT of noisy speech and estimated noise can be calculated simultaneously, and an inverse FFT (IFFT) will then be utilized to estimate speech signal. The computation of EVD (SVD) for the 5×5 matrix costs slightly less (more) time than that of 256-point FFT plus IFFT. The proposed approaches are of the same computational complexity as the original SS or MF methods.

The proposed algorithm transforms a speech enhancement problem into a noise extraction problem. It is very hard to extract a desired signal in the additive noise model. Fortunately, the noise signal is strong when the speech is inactive. Hence, we can extract the additive noise in the associated inactive frames with high correlation. The corresponding performances measured by output SNR and SD show its superiority over conventional approaches. Therefore the MSCPE extraction algorithm can be regarded as *attractive*, and the feasibility of noise extraction from noisy speech is convinced.

As for the artifacts, conventional SS and WF approaches usually bring music noises, the same as the modified SS and WF approaches. The pdf-based speech enhancement algorithms, such as MMSE- or MAP-based ones, may help to reduce the music noise. In our further work, the proposed two-stage idea will be incorporated into pdf-based algorithms to remedy this problem.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] C. He and G. Zweig, "Adaptive two-band spectral subtraction with multi-window spectral estimation," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Phoenix, AZ, 1999, vol. 2, pp. 793–796.

[3] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Orlando, FL, May 2002, vol. 4, pp. 4164–4167.

[4] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[5] I. Y. Soon and S. N. Koh, "Low distortion speech enhancement," *Proc. Inst. Elec. Eng.*, vol. 147, no. 3, pp. 247–253, 2000.

[6] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[7] C. Breithaupt and R. Martin, "MMSE estimation of magnitude-squared DFT coefficients with super-Gaussian priors," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, China, Apr. 2003, vol. 1, pp. 896–899.

[8] A. Hyvarinen, "Sparse code shrinkage: Denoising of nonGaussian data by maximum likelihood estimation," *Neural Computat.*, vol. 11, no. 7, pp. 1739–1768, 1999.

[9] J.-H. Lee, H.-Y. Jung, T.-W. Lee, and S.-Y. Lee, "Speech enhancement with MAP estimation and ICA-based speech features," *Electron. Lett.*, vol. 36, pp. 1506–1507, 2000.

[10] T. Letter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 1110–1126, 2005.

[11] X. Zou, P. Jancovic, J. Liu, and M. Kokuer, "Speech signal enhancement based on MAP algorithm in the ICA space," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1812–1820, May 2008.

[12] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.

[13] Y. Ephraim, Y. Ephraim, and H. Lev-Ari, "A brief survey of speech enhancement," in *The Electronic Handbook*. Boca Raton, FL: CRC, April 2005.

[14] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian-Gaussian mixture," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 896–904, Sep. 2005.

[15] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 87–95, Feb. 2001.

[16] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[17] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Trans. Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[18] G. Wang, N. Rao, S. Shepherd, and C. Beggs, "Extraction of desired signal based on AR model with its application to atrial activity estimation in atrial fibrillation," *EUROSIP J. Adv. Signal Process.*, 2008, Art. No. 728409, 9 pp.

[19] A. Hyvärinen, J. Karhunen, and E. Oja, *Indepedent Component Analysis*. New York: Wiley, 2003.

[20] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2003.

[21] A. Cichocki, R. Thawonmas, and S. Amari, "Sequential blind signal extraction in order specified by stochastic properties," *Electron. Lett.*, vol. 33, no. 1, pp. 64–65, 1997.

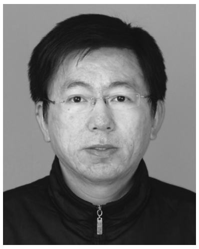
[22] A. K. Barros and A. Cichocki, "Extraction of specific signals with temporal structure," *Neural Comput.*, vol. 13, no. 9, pp. 1995–2000, 2001.

[23] W. Liu, D. P. Mandic, and A. Cichocki, "A class of blind source extraction algorithms based on a linear prediction," *IET Signal Process.*, vol. 1, no. 1, pp. 29–34, 2007.

[24] W. Liu, D. P. Mandic, and A. Cichocki, "Blind second-order source extraction of instantaneous noisy mixtures," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 9, pp. 931–935, Sep. 2006.

[25] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[26] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588–601, 2007.



Gang Wang received the B.E. degree in communication engineering and the Ph.D. degree in biomedical engineering from the University of Electronic Science and Technology of China, Chengdu, in 1999 and 2008, respectively.

He is currently a Lecturer, Institute of Intelligent Systems and Information Technology, School of Electronic Engineering, University of Electronic Science and Technology of China. His current research interests include blind signal processing and intelligent systems.

Chunguang Li received the M.S. degree in pattern recognition and intelligent systems and the Ph.D. degree in circuits and systems from the University of Electronic Science and Technology of China, Chengdu, in 2002 and 2004, respectively.

Currently, he is a Professor with the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. His current research interests include computational neuroscience, statistical signal processing, and computer vision and audio.

Le Dong received the B.E. degree in telecommunication engineering and the M.E. degree in communications and information systems from Xidian University, Xi'an, China, in 2001 and 2004, respectively, and the Ph.D. degree in electronic engineering from Queen Mary, University of London, London, U.K., in 2009.

She is currently the Associate Professor of Institute of Intelligent Systems and Information Technology, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu. Her current research interests include multimedia fusion, machine vision, knowledge representation, intelligent systems, and artificial intelligence.