Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 509–522

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Real and imaginary modulation spectral subtraction for speech enhancement

Yi Zhang, Yunxin Zhao *

*Department of Computer Science, University of Missouri-Columbia, Columbia, MO 65211, USA*

Received 11 October 2011; received in revised form 23 August 2012; accepted 5 September 2012
Available online 6 November 2012

**Abstract**

In this paper, we propose a novel spectral subtraction method for noisy speech enhancement. Instead of taking the conventional approach of carrying out subtraction on the magnitude spectrum in the acoustic frequency domain, we propose to perform subtraction on the real and imaginary spectra separately in the modulation frequency domain, where the method is referred to as MRISS. By doing so, we are able to enhance magnitude as well as phase through spectral subtraction. We conducted objective and subjective evaluation experiments to compare the performance of the proposed MRISS method with three existing methods, including modulation frequency domain magnitude spectral subtraction (MSS), nonlinear spectral subtraction (NSS), and minimum mean square error estimation (MMSE). The objective evaluation used the criteria of segmental signal-to-noise ratio (Segmental SNR), PESQ, and average Itakura–Saito spectral distance (ISD). The subjective evaluation used a mean preference score with 14 participants. Both objective and subjective evaluation results have demonstrated that the proposed method outperformed the three existing speech enhancement methods. A further analysis has shown that the winning performance of the proposed MRISS method comes from improvements in the recovery of both acoustic magnitude and phase spectrum.

*Keywords:* Spectral subtraction; Noise reduction; Speech phase; Modulation frequency

## 1. Introduction

The goal of speech enhancement is to improve speech quality in noisy environment, which requires finding a good tradeoff between noise reduction and speech distortion introduced during the enhancement process. Various speech enhancement techniques have been generated and applied to the real world noisy speech. In general, by using more hardware to acquire spatial information of a target speech source, multi-channel speech enhancement techniques (Farrrel et al., 1992; Yellin and Weinstein, 1996) can provide enhancement performance superior to single channel enhancement methods. However, due to its convenient implementations, single channel speech enhancement has remained a hot spot in speech research. Some widely used single channel speech enhancement methods include spectral subtraction, Wiener filtering, and MMSE, etc.

Spectral subtraction is one of the most widely used speech enhancement techniques (Boll, 1979). Spectral subtraction methods typically focus on signal magnitude spectrum and use noisy phase spectrum in signal reconstruction, where the signal magnitude spectrum is estimated by subtracting an estimate of the noise magnitude spectrum from the noisy signal magnitude spectrum. A major drawback of the spectral subtraction approach is the introduced musical tone in the enhanced speech which is caused by the mismatch of the noise estimate and the true noise.

Wiener filtering (Chen et al., 2006) aims at reducing noise by minimizing the mean square error between the estimated and the clean speech signals. The major shortcoming of the Wiener filter approach is the requirement of a priori knowledge of the power spectrum of the clean speech.

* Corresponding author. Tel.: +1 573 882 3374; fax: +1 573 882 8318.
*E-mail addresses:* yzcb3@mail.missouri.edu (Y. Zhang), Zhaoy@missouri.edu (Y. Zhao).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.09.005

MMSE (Ephraim and Malah, 1984) uses a Bayesian approach to determine the clean speech magnitude spectrum assuming Gaussian distributions for the speech and noise magnitudes. It is worth noting that under this assumption, noisy speech phase was proved to be the optimal phase for the enhanced speech, and hence only the magnitude MMSE has been used in speech enhancement applications.

Speech phase spectrum has been considered insignificant in perceptual speech quality (Wang and Lim, 1982), and so traditional enhancement methods focus on magnitude spectrum enhancement and use noisy phase spectrum in reconstructing speech. When SNR is high, noisy speech phase is indeed close to clean speech phase, and using noisy phase to replace clean phase would not introduce perceptual distortion. However, when SNR drops low, noisy phase plays a more apparent role in the enhanced speech. It has been indicated that when the spectral SNR is lower than approximately 8 dB for all frequencies, a mismatch in phase might be perceived as "roughness" in speech quality (Loizou, 2007), which means that under this condition, even if we had the exact clean speech magnitude spectrum, we would not be able to recover the clean speech signal with unperceivable distortion.

Recently, more interests in speech phase have been reported. Phase information was used to generate features for automatic speech recognition (Schluter and Ney, 2001; Zhu and Paliwal, 2004; Hegde and Murthy, 2007), and phase information was applied to improve perceptual quality of enhanced speech. Shannon and Paliwal (2006) investigated estimating the short time Fourier transform (STFT) phase spectrum independently from the STFT magnitude spectrum for speech enhancement applications and observed substantial improvements in noise reduction and speech quality. Wójcicki et al. (2008) proposed phase spectrum compensation to control the amount of reinforcement or cancellation that occurs during the synthesis of the enhanced signal by adding an anti-symmetry function to the noisy speech signal in the frequency domain. Aarabi and Shi (2004) proposed phase-error filtering based on the assumption that phase variations between multiple microphone channels after time delay compensation are due purely to the influence of the background noise, where the observed between-channel phase difference is used to filter noisy speech such that a larger phase difference results in a greater signal attenuation. Lu and Loizou (2008) proposed a geometric spectral subtraction approach that addressed the shortcomings of spectral subtraction concerning musical noise and speech-noise cross-term issues, where they used the phase differences between the noisy signal and the noise to estimate the cross-terms. Fardkhaleghi and Savoji (2010) investigated the role of phase spectrum in speech enhancement using Wiener filtering and minimum statistics and showed that better results are achieved using phase correction for different noise types. Kleinschmidt et al. (2011) proposed a novel method for acquiring phase information and used the phase information to complement the traditional magnitude-only spectral subtraction in speech enhancement, and they obtained good results in a 15–20 dB SNR environment.

In this current work, we propose a new approach to spectral subtraction for enhancing speech signal from noise, where the subtraction processing is performed on the real and imaginary spectra separately, and the separately enhanced spectra are used to recover the complex signal spectra. This approach is supported by our experimental observation that the real, imaginary, and magnitude spectra have similar time–frequency (T–F) characteristics. We carry out the subtraction processing in the modulation frequency domain for the purpose of reducing musical noise as proposed in (Paliwal et al., 2010). Differing from Paliwal et al. (2010) where the noisy speech acoustic magnitude spectra that contain the cross-terms of speech and noise were transformed to the modulation frequency domain for spectral subtraction, our separate transformation of the real and imaginary acoustic spectra to the modulation frequency domain does not carry the acoustic-domain speech-noise cross-terms. Furthermore, unlike many speech enhancement methods, our synthesis of speech signal from the modified acoustic spectra does not use the acoustic phase spectra of the noisy speech. We conducted comparative experiments using the criteria of segmental signal-to-noise ratio (SNR), PESQ, and ISD to evaluate the performance of the proposed MRISS method against three existing methods, including modulation-frequency domain spectral subtraction (MSS), nonlinear spectral subtraction (NSS), and minimum mean-square error (MMSE) estimator, in noisy conditions of five noise types (white, bable, pink, volvo, factory2 noises) and four SNR levels (−5, 0, 5, and 10 dB).

The organization of this paper is as follows. In Section 2, we discuss the background of conventional spectral subtraction algorithms; in Section 3 we introduce our proposed speech enhancement method; in Section 4 we present experimental results, and in Section 5 we give a conclusion.

## 2. Background of spectral subtraction

### 2.1. Acoustic domain spectral subtraction

A typical method of spectral subtraction performed in the acoustic frequency domain is the generalized frame-by-frame subtraction (Berouti et al., 1979; Boll, 1979) defined as:

$$|\widehat{S}(k,t)|^{\gamma}=\begin{cases}|X(k,t)|^{\gamma}-\alpha(k)|\widehat{N}(k,t)|^{\gamma} & if\,|X(k,t)|^{\gamma}>(\alpha(k)+\beta)|\widehat{N}(k,t)|^{\gamma}\\ \beta|\widehat{N}(k,t)|^{\gamma} & otherwise\end{cases}\tag{1}$$

where $|X(k,t)|$ is the noisy speech magnitude spectrum, $|\widehat{N}(k,t)|$ is the noise magnitude spectral estimate, $|\widehat{S}(k,t)|$ is the reconstructed speech magnitude spectrum, $k$ and $t$ are the frequency and the time indices, respectively; $\alpha(k)$ is an over-subtraction factor which is a function of segmental SNR (Kamath and Loizou, 2002), $\beta$ is a spectral flooring factor that controls the effect of over-subtraction and avoids negative spectrum, and $\gamma$ determines the type of spectrum that the subtraction is operated on, i.e., magnitude spectrum if $\gamma = 1$ and power spectrum if $\gamma = 2$.

In general, three kinds of errors are introduced in the conventional magnitude spectral subtraction, consisting of (1) error in noise estimation; (2) error caused by ignoring the speech-noise cross-term in magnitude spectrum; (3) error caused by using noisy phase spectrum with enhanced magnitude spectrum in signal reconstruction.

These errors degrade the performance of speech enhancement. The first type of error had been widely studied, and several techniques (Lin et al., 2003; Martin, 1994; Hirsch and Ehrlicher, 1995) have been developed to track noise efficiently. When SNR is high, the cross-term is relatively small, and noisy phase is close to the phase of clean signal, and thus conventional spectral subtraction methods do not suffer from these two types of errors. However, as SNR decreases, both the cross-term error and the noisy phase error become nonnegligible in signal reconstruction. Some efforts have been reported to address these two types of errors in speech recognition and speech enhancement (Yoma et al., 1998; Kitaoka and Nakagawa, 2004; Evans et al., 2006; Lu and Loizou, 2008).

### 2.2. Modulation domain spectral subtraction

As we mentioned above, a major problem of spectral subtraction is the musical noise introduced by the subtractive processing in the low energy regions of speech. Generally, musical noises can be reduced in several ways (Huang and Benesty, 2004): (1) a careful design of the analysis and synthesis filter bank; (2) a proper use of time averaging techniques in conjunction with appropriate decision criteria; (3) augmenting the traditional gain function with a soft-decision voice activity detection statistic.

Recently Paliwal et al. (2010) indicated that a good compromise between musical noise and temporal slurring could be obtained by performing spectral subtraction in the modulation frequency domain and carefully choosing the modulation frame length (180–280 ms), and thus reducing musical noise typically associated with acoustic frequency domain spectral subtractive algorithms. This method is referred to as MSS and is contrasted with the proposed MRISS in the next section.

## 3. The real-imaginary modulation spectral subtraction

### 3.1. The proposed method

Our proposed spectral subtraction algorithm is described in the block diagram of Fig. 1.

The noisy speech $x(n)$ is first windowed by a Hamming window function $w(n)$ into overlapped frames and each frame is then transformed into the acoustic frequency domain via a $M$-point fast Fourier transform (FFT) to produce the complex spectra

$$X(n,k) = \sum_{l=0}^{M-1} x(l+nP)w(l)e^{-j2\pi lk/M} \tag{2}$$

where $k = 0, 1, \ldots, M-1$ is the frequency index, $n$ is the time index of the windowed frames, $M$ is the window length, and $P$ is the window shift.

For each acoustic frequency bin, the real and imaginary spectrum $X_R(n, k)$ and $X_I(n, k)$ are again first windowed by a Hamming window function $v(n)$ across time into overlapped time frames, and each frame is then transformed into the modulation frequency domain via an $N$-point FFT

$$Z(k,t,m) = \sum_{n=0}^{N-1} X(n+tD,k)v(n)e^{-j2\pi nm/N} \tag{3}$$

where $m = 0, 1, \ldots, N-1$ is the modulation frequency index, $k$ is the acoustic frequency index, $t$ is the time index, $N$ is the window length, and $D$ is the window shift.

To facilitate spectral subtraction, we consider the noise estimation algorithm of Martin (2001), where the power spectral density of nonstationary noise is estimated from noisy speech signal without using explicit voice activity detection. We apply this estimator to the real and imaginary acoustic spectra to obtain $\widehat{N}_R(n,k)$ and $\widehat{N}_I(n,k)$, and then perform the 2nd FFT transform on $\widehat{N}_R(n,k)$ and $\widehat{N}_I(n,k)$ separately to the modulation domain as described above in Step-2 to obtain $|\widehat{M}_R(k,t,m)|$ and $|\widehat{M}_I(k,t,m)|$, which are used as noise estimates in the subsequent noise subtraction.

In carrying out spectral subtraction, we adopt the magnitude subtraction method proposed by Boll (1979), and extend it into the modulation frequency domain for the separate enhancements of the real and imaginary spectra.

The subtraction computation on the real spectrum is given below in Eq. (4), and that on the imaginary spectrum is defined in a similar way:

$$|\widehat{Z}_R(k,t,m)| = \begin{cases} |Z_R(k,t,m)| - \alpha(t)|\widehat{M}_R(k,t,m)| & if |Z_R(k,t,m)| > (\alpha(t)+\beta)|\widehat{M}_R(k,t,m)| \\ \beta|\widehat{M}_R(k,t,m)| & otherwise \end{cases} \tag{4}$$

where the parameter $\alpha(t) = 2 - \frac{3}{20}SNR(t)$ controls the amount of noise subtraction, the parameter $\beta = 0.005$ controls the spectral floor. The estimated modulation spectra $\widehat{Z}_R(k,t,m)$ is formed by the modified magnitude $|\widehat{Z}_R(k,t,m)|$ and noisy phase $\angle Z_R(k,t,m)$, and in a similar way the $\widehat{Z}_I(k,t,m)$ is formed. The estimated modulation spectra $\widehat{Z}_R(k,t,m)$ and $\widehat{Z}_I(k,t,m)$ are inverse transformed back to the acoustic frequency domain by using the overlap-add method with synthesis windowing to produce $\widehat{X}_R(n,k)$ and $\widehat{X}_I(n,k)$, from which a complex acoustic frequency spectrum $\widehat{X}(n,k)$ is composed. Finally, the time domain speech signal estimate $\hat{s}(n)$ is obtained via the inverse Fourier transform and the overlap-add method.

In the MSS method of Paliwal et al. (2010), the sequence of acoustic magnitude spectra $|X(n, k)|$ was transformed into the modulation frequency domain while the sequence of acoustic phase spectra was untouched. In the modulation

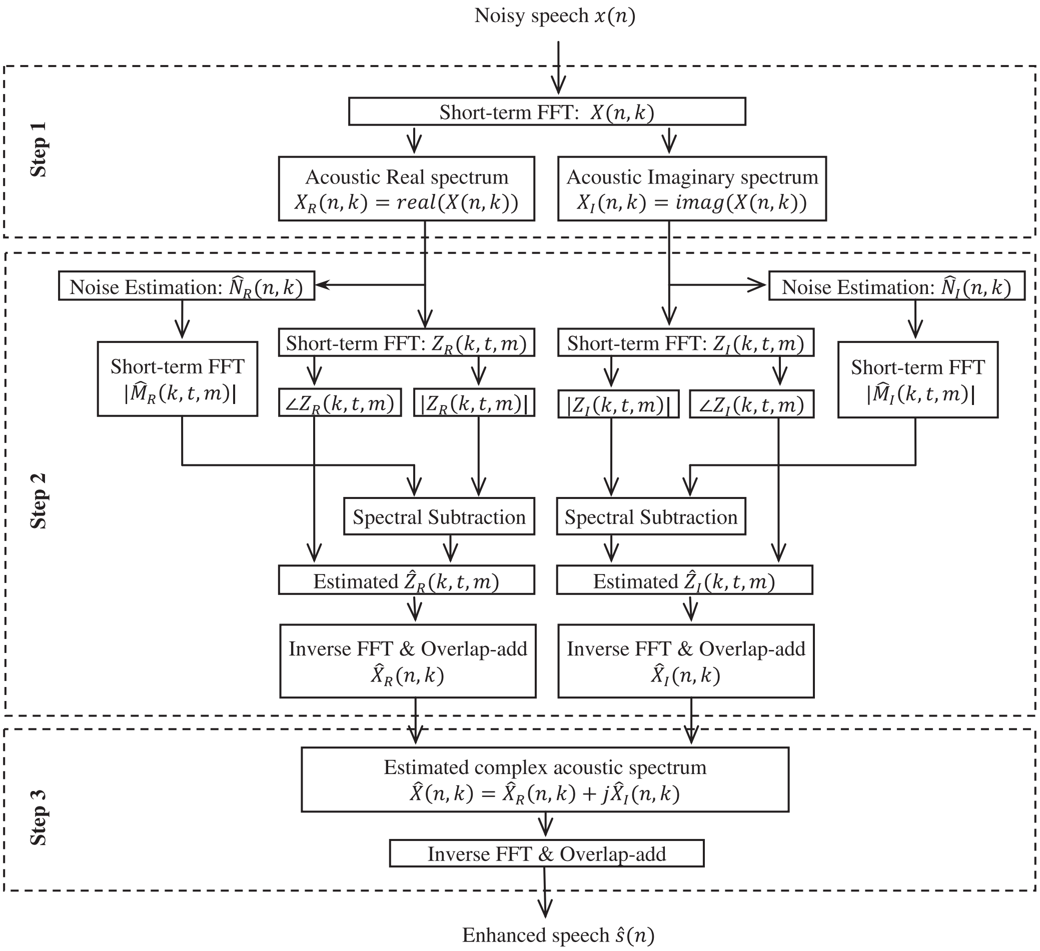

Noisy speech $x(n)$
Step 1
Short-term FFT: $X(n,k)$
Acoustic Real spectrum $X_R(n,k) = real(X(n,k))$
Acoustic Imaginary spectrum $X_I(n,k) = imag(X(n,k))$
Step 2
Noise Estimation: $\hat{N}_R(n,k)$
Noise Estimation: $\hat{N}_I(n,k)$
Short-term FFT $|\hat{M}_R(k,t,m)|$
Short-term FFT: $Z_R(k,t,m)$
Short-term FFT: $Z_I(k,t,m)$
Short-term FFT $|\hat{M}_I(k,t,m)|$
$\angle Z_R(k,t,m)$
$|Z_R(k,t,m)|$
$|Z_I(k,t,m)|$
$\angle Z_I(k,t,m)$
Spectral Subtraction
Spectral Subtraction
Estimated $\hat{Z}_R(k,t,m)$
Estimated $\hat{Z}_I(k,t,m)$
Inverse FFT & Overlap-add $\hat{X}_R(n,k)$
Inverse FFT & Overlap-add $\hat{X}_I(n,k)$
Step 3
Estimated complex acoustic spectrum $\hat{X}(n,k) = \hat{X}_R(n,k) + j\hat{X}_I(n,k)$
Inverse FFT & Overlap-add
Enhanced speech $\hat{s}(n)$

Fig. 1. Block diagram of the proposed method.

frequency domain, a noise estimate was subtracted from the noisy speech magnitude spectra, and the modified speech magnitude spectra coupled with the noisy modulation phase spectra was then transformed back to the acoustic domain. The enhanced acoustic magnitude spectra and the noisy acoustic phase spectra together were transformed back to the time domain to produce the enhanced speech signal.

### 3.2. Properties of the proposed method

Based on the algorithm description of Fig. 1, several properties of our proposed MRISS method are apparently different from conventional spectral subtraction methods. The differences pertain to speech-noise cross-terms, modulation domain spectral subtraction, and the handling of phase spectra in speech signal reconstruction. These three aspects are discussed below.

(1) Speech-noise cross-term in the acoustic frequency domain

For a speech signal corrupted by an additive noise, i.e., $X(n,k) = S(n,k) + N(n,k)$, the squared magnitude spectrum is given as

$$|X(n,k)|^2 = |S(n,k)|^2 + |N(n,k)|^2 + 2|S(n,k)||N(n,k)|\cos(\theta_\Delta(n,k)) \quad (5)$$

where $k$ and $n$ are the frequency and time indices, $\theta_\Delta(n,k) = \theta_s(n,k) - \theta_n(n,k)$.

By adding and subtracting $2|S(n,k)||N(n,k)|$ on the right hand side of Eq. (5) to complete the square of $(|S(n,k)| + |N(n,k)|)^2$, and then taking square root on both sides, we obtain

$$|X(n,k)| = (|S(n,k)| + |N(n,k)|) \cdot \sqrt{1 + \frac{2\gamma(n,k)}{(1+\gamma(n,k))^2}(\cos(\theta_\Delta(n,k)) - 1)} \quad (6)$$

where $\gamma(n,k) = \frac{|N(n,k)|}{|S(n,k)|}$.

In conventional magnitude spectral subtraction, the speech-noise cross-term $\frac{2\gamma(n,k)}{(1+\gamma(n,k))^2}(\cos(\theta_\Delta(n,k)) - 1)$ is assumed to be zero. This assumption depends on two factors: (1) $\gamma(n,k) \to 0$ or $\gamma(n,k) \to \infty$; (2) $\cos(\theta_\Delta(n,k)) \to 1$.

Fig. 2(a) and (b) show the scatter plots of cross-term vs. SNR (averaged over $\cos(\theta_\Delta(n,k))$) and cross-term vs. $\cos(\theta_\Delta(n,k))$ (averaged over SNR), respectively from a TIMIT speech sentence. It is easily seen that when SNR is far away from 0 dB, the cross-term trends to zero; also,

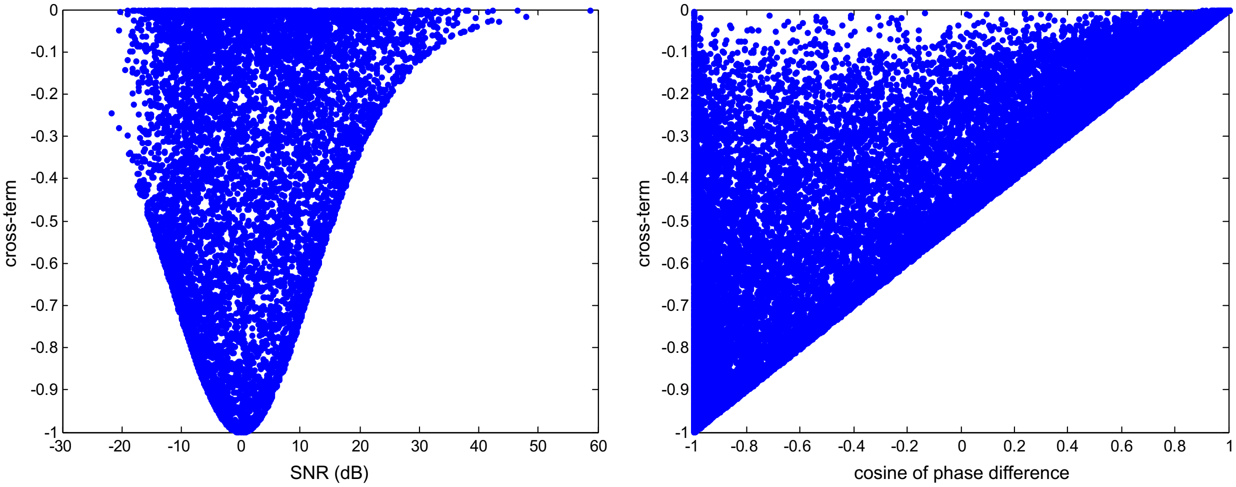


Fig. 2. Relationship between cross-term and (a) SNR, and (b) cosine of phase difference (sumed over all frequency bins).

when $\cos(\theta_{\Delta}(n,k))$ is close to one, the cross-term is close to zero too.

In our proposed MRISS method, as shown in Step 1 of Fig. 1, the real and imaginary spectra are separately transformed into the modulation frequency domain, and therefore the cross-term in $|X(n,k)|$ is avoided. Only in the modulation frequency domain MRISS produces cross-terms in $|X_R(k,t,m)|$ and $|X_I(k,t,m)|$. In contrast, if the magnitude spectrum $|X(n,k)|$ is transformed into the modulation frequency domain as in the method of MSS, then the complex modulation spectra will contain the effect of the acoustic frequency domain cross-terms, and when the magnitude modulation spectra are further computed, additional cross-terms will be produced in the modulation frequency domain.

In Fig. 3, we further compare the distribution of the cross-term (generated from the same sentence with Fig. 2), in the acoustic domain and modulation domain. We observe that the cross-term distribution in the real or imaginary modulation domain is somewhat more concentrated on zero, which means moving the cross-term from acoustic domain to modulation domain at least did not degrade the performance.

(1) Modulation frequency domain spectral enhancement

Denote the complex acoustic spectrum as $X(n,k) = |X(n,k)|\exp\{j\Theta(n,k)\}$, where $\Theta(n,k)$ is the acoustic phase spectrum. When FFT is applied on the time sequence of acoustic magnitude spectrum $|X(n,k)|$ to produce the modulation spectrum as in the MSS method (Paliwal et al., 2010), the resulting modulation spectral energy is concentrated in low modulation frequency since $|X(n,k)|$ varies slowly with time, which is shown in Fig. 4(a) for a fixed harmonic subband $k$. In the MRISS method, FFT is applied separately on the real and imaginary acoustic frequency spectra. For the real acoustic spectra, $X_R(n,k) = |X(n,k)|\cos\{\Theta(n,k)\}$, and so in the modulation domain $Z_R(k,t,m) = Z_M(k,t,m) \circledast Z_{COS}(k,t,m)$, with

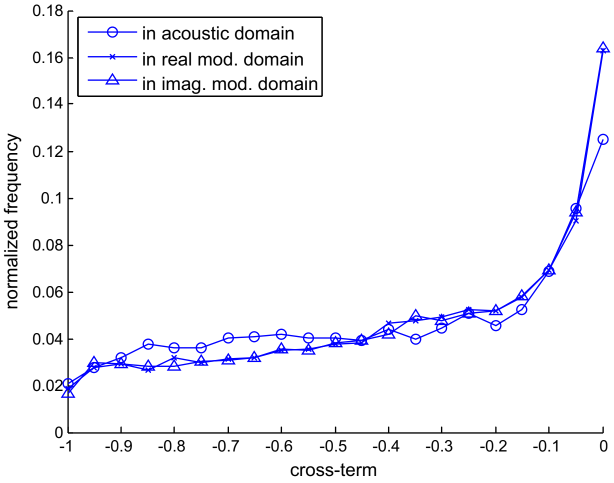


Fig. 3. Histogram of the cross-term in acoustic and modulation domains.

$Z_{COS}(k,t,m) = FFT\{\cos(\omega_k n + \phi_{n,k})\}$, and $\circledast$ denotes convolution in $m$. Similarly, $Z_I(k,t,m) = Z_M(k,t,m) \circledast Z_{SIN}(k,t,m)$. $Z_{COS}(k,t,m)$ and $Z_{SIN}(k,t,m)$ are shown in Fig. 4(b) and (d), where we can see that in each acoustic frequency subband $k$, $\cos\{\Theta(n,k)\}$ and $\sin\{\Theta(n,k)\}$ are quasi-sinusoidal signals (with limited bandwidth) and the frequency components vary with time $n$, which reflects the speech frequency variation. Compared with MSS, $Z_R(k,t,m)$ is a convolution of the $Z_M(k,t,m)$ in Fig. 4(a) with $Z_{COS}(k,t,m)$ in Fig. 4(b), which spreads the signal energy distribution in the modulation spectra, as shown in Fig. 3(c).

The different characteristics in the modulation spectra of $Z_M(k,t,m)$, $Z_R(k,t,m)$ and $Z_I(k,t,m)$ thus have different impacts on the spectral subtraction outcomes of MSS and MRISS.

(1) Phase recovery in acoustic frequency domain

The instantaneous phase of a complex signal $u(t)$ is $\varnothing(t) = \arg(u(t))$, a function of the real and imaginary

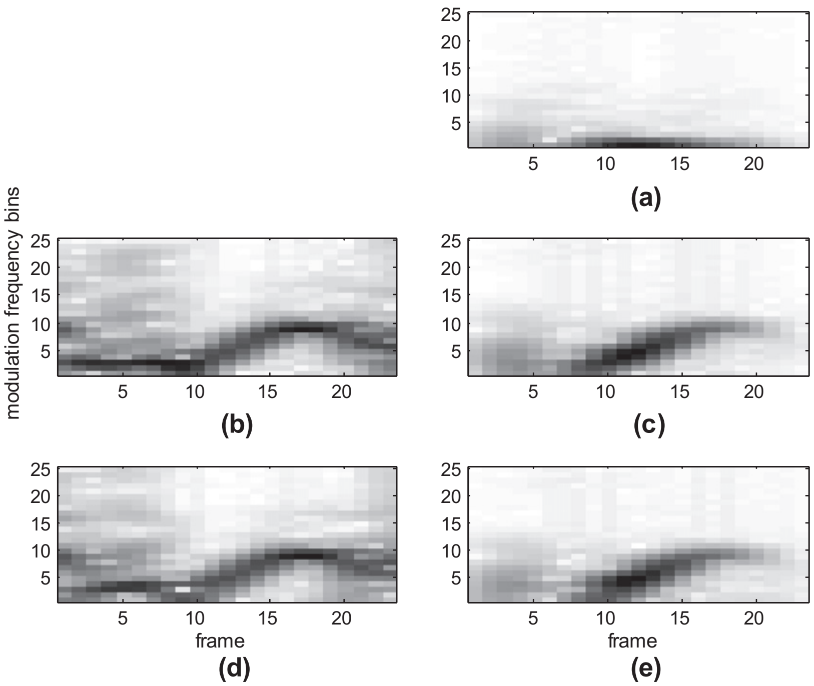


Fig. 4. Modulation spectra of one acoustic frequency subband (a) $Z_M(k,t,m)$, (b) $Z_{COS}(k,t,m)$, (c) $Z_R(k,t,m)$, (d) $Z_{SIN}(k,t,m)$ and (e) $Z_I(k,t,m)$.

components of $u(t)$. The energy of voiced speech concentrates on its harmonics, where the harmonic subband signals are each sinusoidal-like with structured phase. This characteristic of voiced speech is reflected in the narrowly peaked distribution of the temporal difference of the instantaneous phase in each speech harmonic subband signal, defined here as $\Delta\varnothing(t) = \varnothing(t) - \varnothing(t-1)$ with $t$ indexing speech frames. In contrast, wide band noises such as white, babble, pink noises have random phase and thus random instantaneous phase difference. Fig. 5 shows the histogram of $\Delta\varnothing(t)$ computed from an isolated vowel /a/ in a speech harmonic subband (centered at 600 Hz, with a 16 Hz bandwidth), and the histogram of $\Delta\varnothing(t)$ of white noise of the same subband (the two subband signals were both 1.6 s long, the analysis window length was 25 ms, and the window shift was 2.5 ms). As expected, the distribution of the speech instantaneous phase difference has a sharp peak while that of the white noise is broad. From this perspective, voiced speech phase can be enhanced through denoising the real and imaginary components of the speech harmonic structure, and the obtained acoustic complex spectra can then be used in speech signal recovery.

To illustrate the effect of the modulation-domain real-imaginary spectral processing on speech phase recovery, Fig. 6 compare the modulation spectra $Z_M(k,t,m)$ and $Z_R(k,t,m)$ of the above two subband signals (the vowel /a/ and the white noise at SNR 5 dB). As in Fig. 4, the energy of $Z_M(k,t,m)$, for either speech or noise, concentrates in low frequency; in contrast, the energy of $Z_R(k,t,m)$ of speech concentrates in a narrow, time-varying mid band, while that of the white noise spreads out. This suggests that the energies of speech and noise overlap less in $Z_R(k,t,m)$ than in $Z_M(k,t,m)$. Therefore, for speech harmonics where SNR is higher than other spectral regions, the SNR is further improved in $Z_R(k,t,m)$.

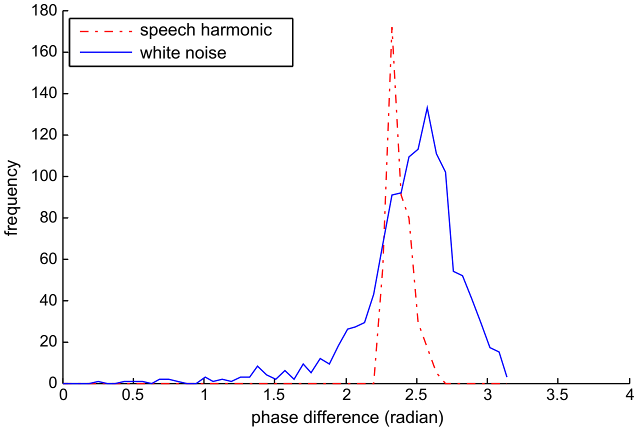


Fig. 5. Histograms of instantaneous phase difference of voiced speech and white noise.

To measure the difference in energy distributions of speech and noise corresponding to $Z_R(k,t,m)$ of Fig. 6, $|Z_R(k,t,m)|$ is normalized by $\sum_{t,m}|Z_R(k,t,m)|$ to become a probability distribution over $(t,m)$, and such a normalized distribution of speech is referred to as $S_R(k,t,m)$ and that of noise as $N_R(k,t,m)$. Kullback–Leibler (K–L) divergence is then computed for the two distributions as

$$D_{KL}(S_R,N_R) = \sum_t\sum_m S_R(k,t,m) ln \frac{S_R(k,t,m)}{N_R(k,t,m)}$$

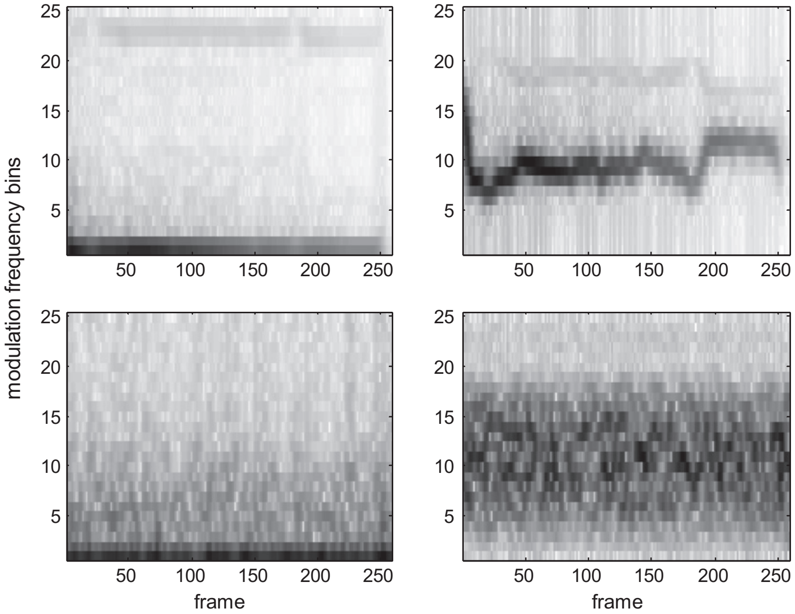


Fig. 6. Modulation spectra of $Z_M(k, t, m)$ (left) and $Z_R(k, t, m)$ (right) of vowel /a/ (top) and white noise (bottom) at the subband 600 Hz.

Since K–L divergence is asymmetric, $D_{KL}(N_R, S_R)$ is also computed. In a similar way, $|Z_M(k, t, m)|$ is normalized for speech and noise, respectively, referred to as $S_M(k, t, m)$ and $N_M(k, t, m)$, and from the two distributions $D_{KL}(S_M, N_M)$ and $D_{KL}(N_M, S_M)$ are computed. The measured divergence values are 1.31 and 0.66 for $D_{KL}(S_{R,}N_R)$ and $D_{KL}(S_M, N_M)$, and 1.75 and 1.24 for $D_{KL}(N_R, S_R)$ and $D_{KL}(N_M, S_M)$, respectively, confirming less overlap between $S_R$ and $N_R$ than that between $S_M$ and $N_M$. Since in high SNR regions noisy speech phase is close to clean speech phase and speech magnitude can be well recovered, the acoustic real and imaginary components of the speech harmonics can be recovered, and hence speech phase can be enhanced.

It is worth noting that unvoiced speech has unstructured phase in general, making its phase nondiscriminable from that of noise, and MRISS processing is not targeting at recovering speech phase for this type of speech sounds.

## 4. Experimental evaluations

We first illustrate the effectiveness of the proposed method in signal phase estimation. We then evaluate the performance of the proposed method in enhancing speech under five types of noise conditions with three commonly used criteria. A listening test was also conducted under a simplified setting. The MRISS processing parameters are given in Table 1.

### 4.1. Phase estimation

We investigated signal phase in noise for three tasks. One was to estimate the phase of a sinusoidal signal, another was to estimate the phase of speech vowel, and the last was to estimate the direction of arrival (DOA) of two speech sources from two microphone recordings, which used complex time–frequency representations of the signals from the individual microphone recordings.

The signal phase was estimated by $\angle\hat{\theta}(n,k) = \arctan(\widehat{X}_I(n,k)/\widehat{X}_R(n,k))$. The phase error before and after the enhancement processing was computed as $\Delta\theta(n,k) = \angle\theta_c(n,k) - \angle\theta'(n,k)$, where $\angle\theta_c(n,k)$ is the clean phase and $\angle\theta'(n,k)$ is the noisy or enhanced phase.

(1) Sinusoidal signal

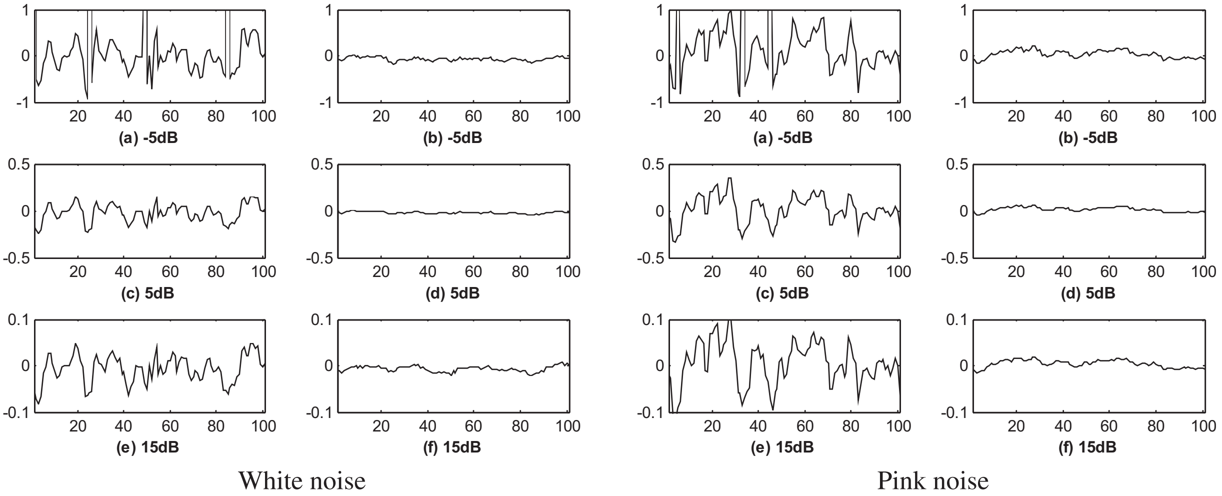
A 50-Hz sinusoidal signal was corrupted by an independent additive noise, producing the noisy signal $x(t) = A{\cdot}\cos(2\pi f_0 t + \theta_0) + n(t)$, where $n(t)$ was white or pink noise with the SNRs ranging from −5 to 15 dB.

Fig. 7 shows the phase errors of a period (100 frames) of the above sinusoidal signal before and after the proposed MRISS for the conditions of white and pink noise, respectively. To avoid crowding the figures, we only show the phase errors at SNRs of −5, 5 and 15 dB. For each noise

Table 1
Experimental parameter setting.

| | Window | Hamming |
|---|---|---|
| Acoustic domain | Window length | 25 ms |
| | Frame shift | 2.5 ms |
| | FFT point | 512 |
| Modulation domain | Window length[1] | 120 ms |
| | Frame shift | 15 ms |
| | FFT point | 32 |

[1] We obtained best results for both MSS and MRISS when we chose modulation window length as 120 ms, instead of 180–256 ms as suggested in Paliwal et al. (2010).



(a) -5dB (b) -5dB (c) 5dB (d) 5dB (e) 15dB (f) 15dB

Fig. 7. Phase errors in white noise (left) and in pink noise (right): (a) (c) (e) before processing; (b) (d) (f) after processing.

type, the left column shows the difference between the noisy and the clean phases, and the right column shows the errors between the estimated and the clean phases. The horizontal axis represents signal sample indices and the vertical axis represents the phase errors.

It is observed that when SNR was high, the noisy phase of $x(t)$ was close to the clean signal phase $\angle\theta_c$, and when SNR was low, the noisy phase of $x(t)$ was very different from the clean phase. The proposed method was able to recover the signal phase well for the sinusoidal signal in both white and pink noises at the different SNR levels.

(1) Speech phase recovery

An isolated vowel (/a/) signal of about 2 s long was corrupted by white and babble noises at SNR of 5 dB, the sampling rate being 8000 Hz. MRISS was performed on the noisy speech's real and imaginary modulation spectra and the recovered acoustic phase was obtained by transforming the modulation spectra back into acoustic domain. We computed the errors of the estimated phase with reference to the clean speech phase, $\Delta\theta(n,k)$, from the time–frequency $(n,k)$ elements with their SNRs in the range of $-5$ to 15 dB. The exclusion of the $(n,k)$ elements outside this SNR range is based on the consideration that when SNR is very low, the speech phase is too noisy to be recovered, and when SNR is very high, the noisy speech phase is already sufficiently close to the clean speech phase. In Fig. 8 we show the histograms of the phase errors thus generated in the two types of noises, and for reference, we also include the histograms of the phase errors from the noisy speech with reference to the clean speech. It is observed that in comparison with the noisy speech phase errors, the errors of the recovered phase are significantly more concentrated around 0, indicating that the recovered phase was closer to the true speech phase in the SNR range of $-5$ to 15 dB, and thus confirming the phase enhancing effect of MRISS.

(1) Direction of Arrival

We consider using a 2-microphone array to estimate the DOA of two simultaneous speech sources. According to the sparsity assumption of speech signals (Yılmaz and Rickard, 2004), a T–F element of the T–F distribution of the mixed speech is dominated by the energy of only one speech source generally and therefore the energy of the two simultaneous sources are distributed in different T–F elements. Expressing the signal arrival time delay $\tau_{12}$ at the two microphones as a function of the sound speed $c$, the microphone spacing $d$, and the arrival angle $\theta_{12}$ leads to

$$\frac{X_1(n,\omega)}{X_2(n,\omega)} \approx \exp\{j\omega\tau_{12}\} = \exp\{j\omega c^{-1} d\cos\theta_{12}\} \qquad (7)$$

where $X_1(n,\omega)$ and $X_2(n,\omega)$ are the complex spectra of the signals acquired by the microphones 1 and 2, respectively, and $\theta_{12}$ is the direction angle of one of the signal sources that has dominant energy at the T–F element $(n,\omega)$ (Araki et al., 2006). From the T–F transforms $X_1(n,\omega)$ and $X_2(n,\omega)$, a histogram is generated by counting the number of T–F elements $(n,\omega)$ that satisfy Eq. (7) for each fixed angle $\theta_{12}$, and the locations of the two largest peaks in the histogram are taken as the DOAs of the speech sources.

For this experiment, a two-source speech mixture was generated by using the anechoic room impulse responses in the RWCP database (RWCP Sound Scene Database in Real Acoustic Environments, 2001) with white and babble noises added to a speech mixture at 0 dB SNR, where the inter-microphone distance was 5.85 cm and the speaker-to-microphone distance was about 2 m. Fig. 9 shows the experiment setup, where $\theta_{12}$ in Eq. (7) is either $\angle\theta_1$ or $\angle\theta_2$.

From Eq. (7), we can see that if frequency $\omega$ is very low, then the phase difference obtained between two microphone inputs is insignificant; on the other hand, if frequency $\omega$ is very high, then phase wrapping is needed to confine phase in the range of $[-\pi,\pi]$. In order to obtain

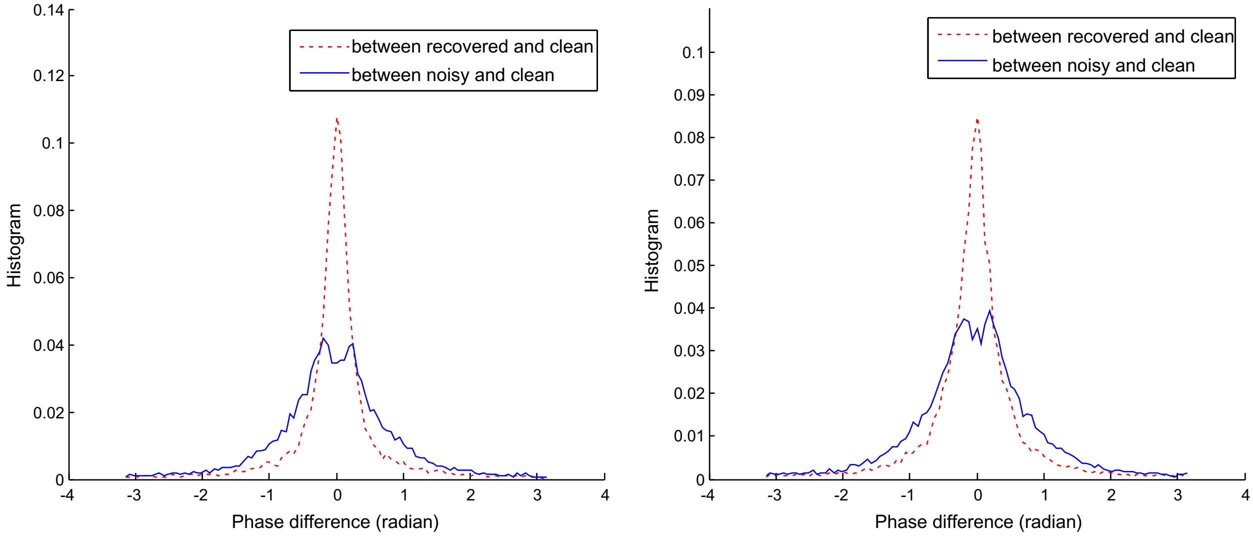


Fig. 8. Histograms of phase errors in white (left) and babble (right) noises within the SNR ranges of −5 to 15 dB.

a good resolution in the DOA histogram and to avoid the need for phase wrapping, a subband of frequency bins (from 2.5 k to 2.9 k Hz) was used to derive each DOA histograms from a block of 2.25 s speech (36,000 samples) that corresponds to around 70 512-point FFT analysis frames. The histograms before and after the proposed processing are shown in Fig. 10. Without the proposed enhancement processing, the DOA histograms (top) could not show two source directions, while after the processing, the DOA histograms (bottom) each showed two peaks clearly, from which one could easily distinguish the two source directions (the dotted lines represent the true source directions). The proposed method therefore holds a good potential of significantly improving DOA estimation of multiple speech sources to enable speech source separation in noisy environments.

### 4.2. Speech enhancement

We evaluated the speech enhancement performances of the proposed method using both subjective and objective measures. Objective measures include the segmental SNR, PESQ, and average Itakura–Saito spectral distance. The results were compared against three existing methods: MSS (Paliwal et al., 2010), NSS (Zhu and Alwan, 2002), and MMSE (Ephraim and Malah, 1984). These three methods were chosen as comparison benchmarks since MSS applies magnitude spectral subtraction in modulation domain, NSS indirectly uses phase information in acoustic domain spectral subtraction, and MMSE is a commonly used method for speech enhancement.

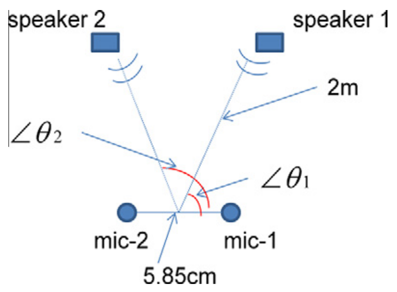


Fig. 9. DOA experiment setup.

We used 40 sentences from the TIMIT dataset as the clean speech. The 40 sentences came from two male and two female speakers, and each speaker contributed 10 sentences. The clean speech was corrupted by five types of noises in the NoiseX92 database, consisting of white, babble, pink, car_volvo, and factory2 noises, and the noisy speech was sampled at 8000 Hz. In these four methods, the same noise estimation algorithm in (Martin, 2001) was used to keep all methods on the same baseline. In NSS and MMSE, the noise estimation was implemented on acoustic magnitude spectrum, while in MSS and MRISS, the noise estimation was implemented on modulation magnitude spectrum.

For every measure criterion and noise type, our proposed method delivered the best performance in almost all SNR conditions, as detailed below in the evaluation experiments Section 4.2.1–4.2.4 We therefore conducted a statistical significance test on the performance difference between the proposed method (best) and the second best performing method in the evaluation Section 4.2.1–4.2.3 experiments where the difference was assumed to be a Gaussian random variable with an unknown variance, and the significance test was one-sided student-$t$ test with $n-1=39$ degree of freedom at the significance level of $\alpha=0.05$ ($t_\alpha=1.686$) Papoulis, 1991.

#### 4.2.1. Segmental signal-to-noise ratio (Segmental SNR)

Segmental SNR is defined as the average SNR values calculated from short segments of speech

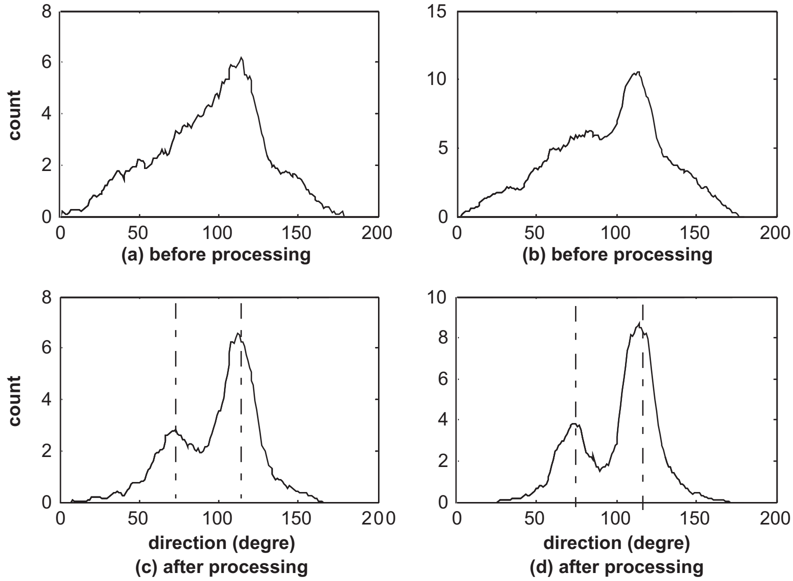


Fig. 10. DOA histogram (left: white noise, right: babble noise).

$$\mathrm{SegSNR} = \frac{1}{N}\sum_{n=0}^{N-1} 10\log_{10} \sum_{k=0}^{K-1} \frac{|s(n,k)|^2}{|s(n,k) - \hat{s}(n,k)|^2}$$

in which $k$ is the frequency index and $n$ is the segment index. In computing the SegSNR values, the segment length was set to be 32 ms (512-point FFT). The larger the segmental SNR value, the better the recovery performance.

From Table 2, we observe that the proposed method provided the largest Segmental SNR in every case, and MSS was always the second best in all the cases. For all five noises and all SNR levels, our proposed method significantly outperformed the MSS method.

### 4.2.2. Perceptual evaluation of speech quality (PESQ)

PESQ is widely adopted for automated assessment of speech quality as experienced by a listener, and a higher PESQ value indicates a better speech quality. We used the PESQ routine of http://www.utdallas.edu/~loizou/speech/software.htm. (Lu and Loizou, 2008) in the experimental evaluation and the results are shown in Table 3.

The proposed method delivered the best performance in most cases. For conditions of white, babble and pink noises, the proposed method outperformed the MSS, NSS and MMSE. For the case of Volvo noise, MMSE delivered the best result at SNR 15 dB, but note that the baseline is already extremely high in this case. For the factory2 noise, only at 15 dB MRISS dropped below MMSE and MSS, and in this case the PESQ for the three methods, including that of the baseline, were all high. It is noted that under the Volvo noise conditions, the differences in PESQ scores among the four methods were small since the base PESQ scores were high. For white, babble, pink and factory 2 noises, when SNR was in the range of –5 to 5 dB, the proposed method significantly outperformed the MSS. The improvement in the volvo noise was not significant due to the high baseline.

Table 2
Comparison on Segmental SNR (dB).

| Input overall SNR | | Noisy | NSS | MMSE | MSS | MRISS |
|---|---|---|---|---|---|---|
| White | −5 | −7.48 | 0.49 | −0.42 | 1.61 | 2.28 |
| | 0 | −3.23 | 3.84 | 3.31 | 4.68 | 5.43 |
| | 5 | 0.71 | 6.85 | 6.81 | 7.59 | 8.04 |
| | 10 | 5.68 | 10.71 | 10.75 | 11.34 | 11.96 |
| | 15 | 8.48 | 14.90 | 14.88 | 15.48 | 16.07 |
| Babble | −5 | −5.30 | −1.00 | −1.07 | −0.62 | 0.33 |
| | 0 | −2.35 | 3.30 | 3.26 | 3.93 | 4.20 |
| | 5 | 0.98 | 5.46 | 5.29 | 6.29 | 6.88 |
| | 10 | 5.34 | 10.35 | 10.27 | 10.74 | 11.06 |
| | 15 | 8.10 | 13.02 | 12.99 | 13.51 | 13.91 |
| Pink | −5 | −7.43 | −0.29 | −0.36 | 0.47 | 1.49 |
| | 0 | −3.19 | 3.23 | 3.00 | 3.73 | 5.05 |
| | 5 | 0.68 | 6.58 | 6.56 | 7.20 | 7.98 |
| | 10 | 5.72 | 10.73 | 10.76 | 11.21 | 12.05 |
| | 15 | 8.51 | 15.17 | 15.10 | 15.77 | 16.45 |
| Volvo | −5 | −7.81 | 6.67 | 6.57 | 8.13 | 9.71 |
| | 0 | −3.49 | 11.47 | 11.51 | 12.18 | 13.66 |
| | 5 | 2.89 | 15.24 | 15.88 | 17.71 | 18.70 |
| | 10 | 7.56 | 20.26 | 20.12 | 20.69 | 21.37 |
| | 15 | 9.38 | 22.15 | 22.03 | 22.76 | 23.13 |
| Factory2 | −5 | −6.30 | 2.74 | 2.37 | 3.51 | 4.74 |
| | 0 | −2.00 | 7.31 | 7.25 | 7.68 | 8.28 |
| | 5 | 2.02 | 10.29 | 10.04 | 11.18 | 12.27 |
| | 10 | 6.00 | 14.84 | 14.78 | 15.64 | 16.55 |
| | 15 | 9.81 | 18.00 | 17.97 | 18.64 | 19.21 |

### 4.2.3. Average Itakura–Saito spectral distance

Itakura–Saito distance (ISD) is a measure of perceptual difference between an original spectrum $P(\omega)$ and an approximation of that spectrum $\widehat{P}(\omega)$, which is defined as:

Table 3
Comparison on PESQ.

| Input overall SNR | | Noisy | NSS | MMSE | MSS | MRISS |
|---|---|---|---|---|---|---|
| White | −5 | 1.56 | 2.15 | 2.17 | 2.27 | 2.39 |
| | 0 | 1.94 | 2.54 | 2.56 | 2.63 | 2.71 |
| | 5 | 2.35 | 2.88 | 2.92 | 2.94 | 2.99 |
| | 10 | 2.66 | 3.22 | 3.25 | 3.25 | 3.29 |
| | 15 | 2.96 | 3.31 | 3.30 | 3.31 | 3.31 |
| Babble | −5 | 1.60 | 2.12 | 2.16 | 2.26 | 2.30 |
| | 0 | 1.77 | 2.33 | 2.40 | 2.52 | 2.58 |
| | 5 | 2.22 | 2.76 | 2.82 | 2.87 | 2.93 |
| | 10 | 2.58 | 3.09 | 3.15 | 3.21 | 3.24 |
| | 15 | 2.91 | 3.28 | 3.27 | 3.30 | 3.30 |
| Pink | −5 | 1.60 | 2.11 | 2.10 | 2.24 | 2.35 |
| | 0 | 1.94 | 2.46 | 2.48 | 2.63 | 2.72 |
| | 5 | 2.35 | 2.81 | 2.79 | 2.90 | 2.96 |
| | 10 | 2.72 | 3.10 | 3.15 | 3.20 | 3.23 |
| | 15 | 2.91 | 3.33 | 3.35 | 3.40 | 3.40 |
| Volvo | −5 | 3.34 | 3.66 | 3.69 | 3.71 | 3.72 |
| | 0 | 3.66 | 3.82 | 3.81 | 3.86 | 3.89 |
| | 5 | 4.00 | 4.12 | 4.13 | 4.12 | 4.15 |
| | 10 | 4.25 | 4.30 | 4.29 | 4.28 | 4.30 |
| | 15 | 4.33 | 4.33 | 4.34 | 4.32 | 4.32 |
| Factory2 | −5 | 2.22 | 2.70 | 2.72 | 2.82 | 2.89 |
| | 0 | 2.64 | 3.12 | 3.15 | 3.22 | 3.26 |
| | 5 | 2.95 | 3.29 | 3.35 | 3.40 | 3.44 |
| | 10 | 3.33 | 3.60 | 3.65 | 3.68 | 3.68 |
| | 15 | 3.65 | 4.10 | 4.12 | 4.12 | 4.10 |

Table 4
Comparison on ISD.

| Input overall SNR | | Noisy | NSS | MMSE | MSS | MRISS |
|---|---|---|---|---|---|---|
| White | −5 | 11.15 | 4.02 | 3.71 | 2.87 | 2.67 |
| | 0 | 7.22 | 3.64 | 3.12 | 2.52 | 2.43 |
| | 5 | 5.64 | 2.16 | 1.96 | 1.80 | 1.74 |
| | 10 | 3.41 | 1.61 | 1.54 | 1.42 | 1.36 |
| | 15 | 2.66 | 1.45 | 1.38 | 1.05 | 0.86 |
| Babble | −5 | 10.20 | 4.66 | 4.57 | 4.42 | 4.37 |
| | 0 | 8.24 | 3.86 | 3.55 | 3.40 | 3.28 |
| | 5 | 6.09 | 2.78 | 2.52 | 2.25 | 2.19 |
| | 10 | 3.74 | 1.47 | 1.27 | 1.07 | 1.02 |
| | 15 | 2.82 | 1.12 | 1.01 | 0.88 | 0.85 |
| Pink | −5 | 10.74 | 4.21 | 3.84 | 3.53 | 3.40 |
| | 0 | 7.75 | 3.55 | 3.30 | 3.05 | 2.92 |
| | 5 | 5.20 | 1.95 | 1.74 | 1.25 | 1.17 |
| | 10 | 3.86 | 1.36 | 1.25 | 1.05 | 1.00 |
| | 15 | 2.65 | 0.98 | 0.90 | 0.74 | 0.69 |
| Volvo | −5 | 4.77 | 1.68 | 1.51 | 1.31 | 1.25 |
| | 0 | 1.67 | 0.80 | 0.78 | 0.72 | 0.71 |
| | 5 | 0.94 | 0.38 | 0.34 | 0.32 | 0.31 |
| | 10 | 0.65 | 0.21 | 0.20 | 0.20 | 0.20 |
| | 15 | 0.25 | 0.18 | 0.17 | 0.18 | 0.17 |
| Factory2 | −5 | 9.54 | 4.30 | 3.86 | 3.43 | 3.23 |
| | 0 | 6.22 | 2.26 | 2.11 | 1.79 | 1.58 |
| | 5 | 4.74 | 1.43 | 1.35 | 1.30 | 1.23 |
| | 10 | 2.85 | 0.75 | 0.66 | 0.56 | 0.45 |
| | 15 | 2.17 | 0.48 | 0.45 | 0.42 | 0.42 |

$$\mathrm{ISD}(P(\omega),\widehat{P}(\omega)) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{P(\omega)}{\widehat{P}(\omega)} - log\frac{P(\omega)}{\widehat{P}(\omega)} - 1\right]d\omega$$

A smaller ISD signifies a higher similarity between the recovered speech and the reference speech. Since the ISD is asymmetric, we used the average ISD instead, which is defined as

$$\textit{averaged}\ \mathrm{ISD}(P1,P2) = (\mathrm{ISD}(P1,P2) + ISD(P2,P1))/2$$

and the averaged ISD is simply referred to as ISD.

As suggested in (Hansen and Pellom, 1998), the largest 5% ISD scores were discarded to exclude the unreliable high distance values. The results are shown in Table 4. Similar with the PESQ test, the Volvo and factory2 noises are less difficult and the ISD scores were lower than the other three noise conditions. The proposed method still obtained the best results across the board. The differences between the proposed method and the second best were significant in the SNR range of –5 to 0 dB for white, babble, pink and factory 2 noises. Similar with the situation of PESQ test, there was no significant improvement in the volvo noise.

### 4.2.4. Subjective evaluation

The subjective evaluation was performed through a sentence-pair listening test. The listening materials included three noise types (white, pink, and babble) at two SNR levels (0, 5 dB) for two speakers (one male and one female), with a total of 12 cases (3*2*2). For each case, one TIMIT speech sentence was used from a speaker (randomly taken from SA1, SA2, and one other sentence in TIMIT) as the dry source, and the three enhancement methods of MMSE, MSS, and MRISS were applied to enhance the speech from noise. The speech sentences enhanced by two different methods were combined pairwisely to generate totally 36 pairs of sentences, from which 4 groups (with overlap) were formed, with each group having 18 sentence pairs and each processing method being used in 12 sentences per group. The play back order of the three methods in each group was balanced, i.e., each of the six combinations of MMSE–MSS, MSS–MMSE, MMSE–MRISS, MRISS–MMSE, MSS–MRISS, MRISS–MSS occurred in three sentence pairs. Listeners were asked to mark one of the three choices for each sentence pair: prefer the first one, prefer the second one, and no preference. Pairwise scoring was employed: a score of +1 was awarded to the preferred method and +0 to the other, and for the no preference response each method was awarded a score of +0.5.

Fourteen normal hearing, native English speakers participated in the experiment. The listening evaluation was conducted in a quiet room. The participants were familiarized with the task during a short practice session before the formal test. Each listener evaluated one of the four groups of sentence pairs. The normalized mean preference score from the subjective evaluation experiment is shown in Fig. 11, where the order of preference is clearly MRISS (0.42), MSS (0.34), and MMSE (0.24). In general, the MRISS processed speech had less residual noise than MMSE, and it introduced less distortion than MSS. The detailed evaluation scores are shown in Table 5, where in each table entry, the first number is the total score that the 1st method was preferred to the 2nd one, the second number is the total score that the 2nd method was

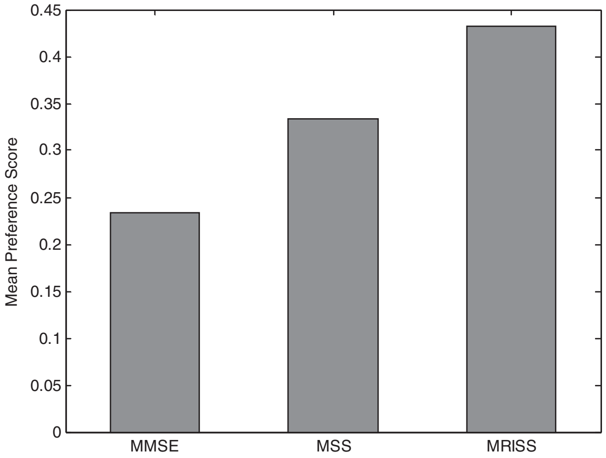


Fig. 11. Subjective evaluation of MMSE, MSS, and MRISS.

Table 5
Comparison on preference score (1st is preferred/2nd is preferred/similar).

| 1st \ 2nd | MMSE | MSS | MRISS |
|---|---|---|---|
| MMSE | – | 30/41/13 | 17/56/11 |
| MSS | – | – | 24/35/25 |
| MRISS | – | – | – |

preferred to the 1st one, and the last number is the total score that the two methods were considered similar.

## 4.3. Performance analysis

We experimentally studied the effects of each of the three factors related to the property of MRISS discussed in Section 3.2. Objective measurements were made in two domains: acoustic frequency domain and time domain. In order to evaluate the performance of modulation domain processing without the confounding factor of acoustic frequency phase, two quality measures on acoustic frequency magnitude spectrum were used, i.e., the ISD and log spectral distance (LSD). In order to evaluate the effect of acoustic frequency phase, we used the measures of PESQ and segmental SNR for the time domain speech signal. The experimental conditions were white, pink and babble noises with the SNRs of −5, 0, 5 and 10 dB.

### 4.3.1. Modulation domain spectral subtraction

In this study, we evaluate the performances of modulation domain magnitude spectral subtractions for the MRISS method and the MSS method. In order to avoid confounding the subtraction evaluation by different use of phase, we set the modulation phase to be the clean speech phase for both MRISS and MSS. For MSS, we evaluated two cases, one used the actual noisy acoustic magnitude spectra which included the speech-noise cross-term, another artificially removed the cross-term.

*4.3.1.1. Case 1: Without cross-term.* In the preprocessing step, we eliminated the cross-term from the acoustic frequency magnitude spectra for MSS (using the known speech and noise data) so that $|\widehat{X}(k,t)| = |S(k,t)| + |N(k,t)|$, and for each fixed k, $|\widehat{X}(k,t)|$ were then transformed to the modulation frequency domain for subtractive enhancement.

*4.3.1.2. Case 2: With cross-term.* In this case, we simply used the noisy acoustic magnitude spectrum $|X(k, t)|$ and transformed it to the modulation frequency domain for subtractive enhancement.

The evaluation results are shown in Fig. 12. We observe that the MRISS method produced better results than the MSS method with or without cross-term, and the fact that the quality of the acoustic frequency magnitude spectra recovered by MSS with the cross-term artificially removed was better than that recovered from the actual magnitude spectrum with the cross-term shows that the cross-term degraded the MSS based enhancement performance.

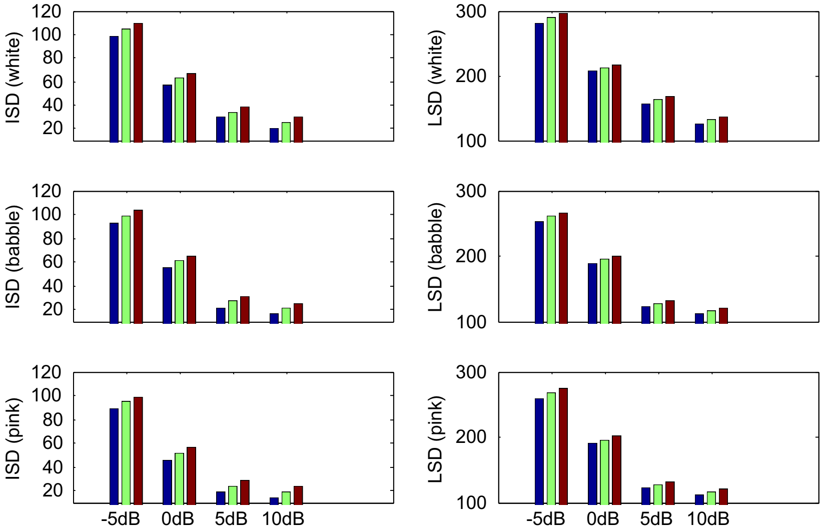


Fig. 12. ISD and LSD evaluations on magnitude recovery (Bars within a SNR group from left to right: MRISS, MSS (without cross term), MSS).

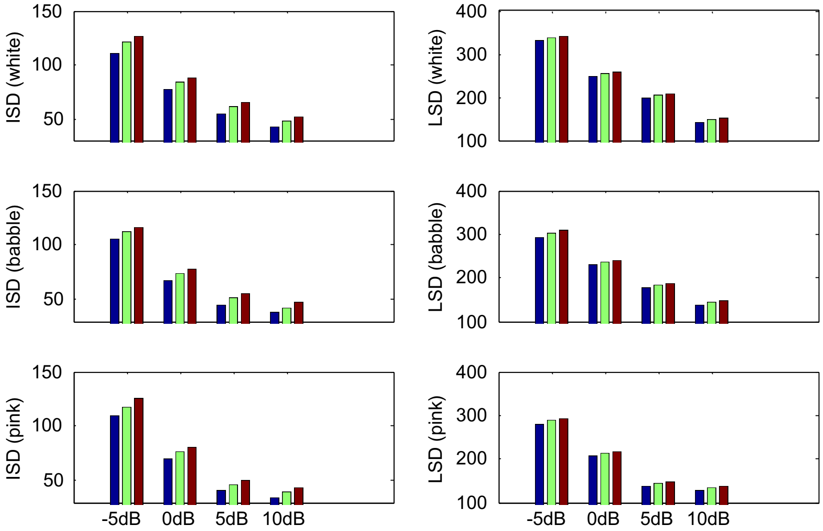


Fig. 13. ISD and LSD evaluations on the modulation domain processing (Bars within a SNR group from left to right: MRISS, MSS (without cross term), MSS).

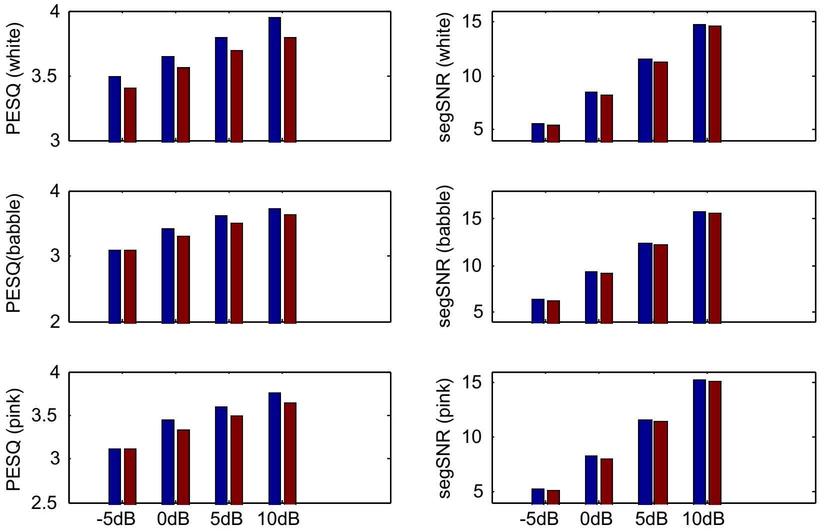


Fig. 14. PESQ and segmental SNR evaluations on the effect of acoustic frequency phase spectra in speech enhancement (Bars within a SNR group from left to right: MRISS, MSS).

*4.3.2. Overall modulation domain processing*

In this study, we evaluated the overall performance of the modulation domain processing, that is, the combination of noisy modulation phase and the spectral subtraction modified modulation magnitude. The evaluation results are shown in Fig. 13. From Fig. 13, we see that the quality of the acoustic magnitude spectra obtained by MRISS is uniformly better than that obtained by the MSS. Both methods showed increased distortion in comparison with Fig. 12 where the clean speech phases were used.

*4.3.3. Acoustic frequency phase spectra*

In this study, we compare the effect of using acoustic frequency phase recovered from the MRISS method against that of using the noisy acoustic frequency phase in the recovered speech signal. We first estimated real and imaginary acoustic spectra using the MRISS method, from which the recovered phase was obtained. We then used the recovered phase and the clean acoustic frequency magnitude spectra to recover the time domain speech signal. For comparison, emulating the MSS method we used noisy acoustic frequency phase spectra and clean acoustic frequency magnitude spectra to recover the time domain speech signal. The results are shown in Fig. 14.

In comparison with using noisy phase, using the MRISS recovered phase obtained an average of 0.1 point gain on PESQ and an average of 0.2 dB gain on segmental SNR over the four SNR and three noise conditions.

## 5. Conclusion

In this paper we have proposed a novel spectral subtraction method for noise reduction in speech. The subtraction is performed in the modulation frequency domain on the real and imaginary spectra separately to preserve the phase information. Our results have shown the capability of the proposed method in estimating signal phase in noise, and in significantly improving the performance of speech enhancement over the existing methods of MSS, NSS and MMSE in most noise and SNR conditions investigated herein (exceptions were in volvo noise and high SNR). A subjective evaluation also showed listeners' preference for our proposed method. Based on our experimental evaluation results, we conclude that both the modulation frequency domain real and imaginary spectra enhancement and acoustic frequency phase spectra contributed to the better quality in the enhanced speech by the MRISS method, where the modulation domain processing played a larger role than the acoustic frequency phase under the studied conditions. The improved acoustic frequency magnitude spectra estimation as well as the enhanced acoustic frequency phase contribute to the superior performance of MRISS over the contrasted spectral subtractive speech enhancement methods. In future work, we shall further investigate the MRISS method's applications in other tasks such as automatic speech recognition and speaker identification.

## References

Aarabi, P., Shi, G., 2004. Phase-based dual-microphone robust speech enhancement. IEEE Trans. Systems Man Cybernet. B: Cybernet. 34, 1763–1773.

Araki, S., Sawada, H., Mukai, R., Makino, S., 2006. DOA estimation for multiple sparse sources with normalized observation vector clustering. In: Proc. IEEE Internat. Conf. on Acoustic Speech Signal Processing, vol. 5, pp. 33–35.

Berouti, M., Schwartz, M., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Internat. Conf. on Acoustic Speech Signal Processing, vol. 23, pp. 208–211.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Chen, J., Benesty, J., Huang, Y., 2006. New insights into the noise reduction wiener filter. IEEE Trans. Acoust. Speech Signal Process. 14, 1218–1234.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32, 1109–1121.

Evans, N., Mason, J., Liu, W., Fauve, B., 2006. An assessment on the fundamental limitations of spectral subtraction. In: Proc. IEEE Internat. Conf. on Acoustic Speech Signal Processing, vol. 1, pp. 145–148.

Fardkhaleghi, P., Savoji, M.H., 2010. New approaches to speech enhancement using phase correction in wiener filtering. Telecommun. (IST), 895–899.

Farrrel, K., Mammone, R.J., Flanagan, J.L., 1992. Beamforming microphone arrays for speech enhancement. In: Proc. IEEE Internat. Conf. on Acoustic Speech Signal Processing, vol. 1, pp. 285–288.

Hansen, J., Pellom, B., 1998. An effective quality evaluation protocol for speech enhancements algorithms. In: Proc. Internat. Conf. on Spoken Language Processing, vol. 7, pp. 2819–2822.

Hegde, R.M., Murthy, H.A., Ramana Rao Gadde, 2007. Significance of the modified group delay feature in speech recognition. IEEE Trans. Audio Speech Lang. Process. 15, 190–202.

Hirsch, H., Ehrlicher, C., 1995. Noise estimation techniques for robust speech recognition. In: Proc. IEEE Internat. Conf. on Acoustic Speech Signal Processing, vol. 23, pp. 153–156.

http://www.utdallas.edu/~loizou/speech/software.htm.

Huang, Y., Benesty, J., 2004. Audio Signal Processing: For Next Generation Multimedia Communication Systems. Kluwer Academic Publishers.

Kamath, S., Loizou, P., 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Proc. IEEE Internat. Conf. on Acoustic Speech, Signal Processing, vol. 200, pp. 4164.

Kitaoka, N., Nakagawa, S., 2004. Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task. In: Proc. Internat. Conf. on Spoken Language Processing, vol. 23, pp. 477–480.

Kleinschmidt, T., Sridharan, S., Manson, M., 2011. The use of phase in complex spectrum subtraction for robust speech recognition. Comput. Speech Lang. 25, 585–600.

Lin, L., Holmes, W., Ambikairajah, E., 2003. Adaptive noise estimation algorithm for speech enhancement. Electron. Lett. 39, 754–755.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press.

Lu, Y., Loizou, P., 2008. A geometric approach to spectral subtraction. Speech Comm. 50, 453–466.

Martin, R., 1994. Spectral subtraction based on minimum statistics. Proc. EUSIPCO, 1182–1185.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process., 504–512.

Paliwal, K.K., Wojcicki, K., Schwerin, B., 2010. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Comm. 52, 450–475.

Papoulis, A., 1991. Probability, Random Variables, and Stochastic Processes, third ed. McGraw-Hill, New York.

RWCP sound scene database in real acoustic environments. 2001. ATR Spoken Language Translation Research Laboratory, Japan.

Schluter, R., Ney, H., 2001. Using phase spectrum information for improved speech recognition performance. In: Proc. IEEE Internat. Conf. on Acoustic Speech Signal Processing, pp. 133–136.

Shannon, B.J., Paliwal, K.K., 2006. Role of phase estimation in speech enhancement. In: Proc. Internat. Conf. on Spoken Language Processing, vol. 23, pp. 1423–1426.

Wang, D., Lim, J., 1982. The unimportance of phase in speech enhancement. IEEE Trans. Acoust. 30, 679–681.

Wójcicki, K., Milacic, M., Stark, A., Lyons, J., Paliwal, K.K., 2008. Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement. IEEE Signal Process. Lett. 15, 461–464.

Yellin, D., Weinstein, E., 1996. Multichannel signal separation: Methods and analysis. IEEE Trans. Acoust. Speech Signal Process. 44, 106–118.

Yılmaz, O., Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. 52, 1830–1847.

Yoma, N., McInnes, F., Jack, M., 1998. Improving performance of spectral subtraction in speech recognition using a model for additive noise. IEEE Trans. Speech Audio Process. 6, 579–582.

Zhu, Q., Alwan, A., 2002. The effect of additive noise on speech amplitude spectra: a quantitative analysis. IEEE Signal Process. Lett. 9, 275–277.

Zhu, D., Paliwal, K.K., 2004. Product of power spectrum and group delay function for speech recognition. In: Proc. IEEE Internat. Conf. on Acoustic Speech Signal Processing, vol. 1, pp. 125–128.