Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 453–466
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A geometric approach to spectral subtraction

Yang Lu, Philipos C. Loizou *

Department of Electrical Engineering, University of Texas-Dallas, Richardson, TX 75083-0688, United States

Received 22 May 2007; received in revised form 18 January 2008; accepted 24 January 2008

Abstract

The traditional power spectral subtraction algorithm is computationally simple to implement but suffers from musical noise distortion. In addition, the subtractive rules are based on incorrect assumptions about the cross terms being zero. A new geometric approach to spectral subtraction is proposed in the present paper that addresses these shortcomings of the spectral subtraction algorithm. A method for estimating the cross terms involving the phase differences between the noisy (and clean) signals and noise is proposed. Analysis of the gain function of the proposed algorithm indicated that it possesses similar properties as the traditional MMSE algorithm. Objective evaluation of the proposed algorithm showed that it performed significantly better than the traditional spectral subtractive algorithm. Informal listening tests revealed that the proposed algorithm had no audible musical noise.

Keywords: Speech enhancement; Spectral subtraction; Musical noise

1. Introduction

The spectral subtraction algorithm is historically one of the first algorithms proposed for noise reduction (Boll, 1979; Weiss et al., 1974), and is perhaps one of the most popular algorithms. It is based on a simple principle. Assuming additive noise, one can obtain an estimate of the clean signal spectrum by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The noise spectrum can be estimated, and updated, during periods when the signal is absent. The enhanced signal is obtained by computing the inverse discrete Fourier transform of the estimated signal spectrum using the phase of the noisy signal. The algorithm is computationally simple as it only involves a single forward and inverse Fourier transform.

The simple subtraction processing comes with a price. The subtraction process needs to be done carefully to avoid any speech distortion. If too much is subtracted, then some speech information might be removed, while if too little is subtracted then much of the interfering noise remains. Many methods have been proposed to alleviate, and in some cases, eliminate some of the speech distortion introduced by the spectral subtraction process (see review in Loizou, 2007, Ch. 5). Some suggested over-subtracting estimates of the noise spectrum and spectral flooring (rather than setting to zero) negative values (Berouti et al., 1979). Others suggested dividing the spectrum into a few contiguous frequency bands and applying different non-linear rules in each band (Kamath and Loizou, 2002; Lockwood and Boudy, 1992). Yet, others suggested using a psychoacoustical model to adjust the over-subtraction parameters so as to render the residual noise inaudible (Virag, 1999).

While the spectral subtraction algorithm can be easily implemented to effectively reduce the noise present in the corrupted signal, it has a few shortcomings. The spectra obtained from the subtractive rules may contain some negative values due to errors in estimating the noise spectrum. The simplest solution is to set the negative values to zero to ensure a non-negative magnitude spectrum. This non-linear processing of the negative values, however, creates small, isolated peaks in the spectrum occurring randomly in time

* Corresponding author. Fax: +1 972 883 4617.
E-mail address: loizou@utdallas.edu (P.C. Loizou).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.01.003

and frequency. In the time-domain, these peaks sound like tones with frequencies that change randomly from frame to frame and introduce a new type of "noise", often called *musical noise* (Berouti et al., 1979). In some cases, the musical noise can be more annoying to the listeners than the original distortions caused by the interfering noise. Other factors contributing to the musical noise phenomenon include the large variance in the estimates of the noisy and noise signal spectra and the large variability in the suppression function.

The derivation of the spectral subtraction equations is based on the assumption that the cross terms involving the phase difference between the clean and noise signals are zero. The cross terms are assumed to be zero because the speech signal is uncorrelated with the interfering noise. While this assumption is generally valid since the speech signal and noise are statistically uncorrelated, it does not hold when applying the spectral subtraction algorithm over short-time (20–30 ms) intervals. Consequently, the resulting equations derived from spectral subtraction are not exact but approximations. Several attempts have been made to take into account or somehow compensate for the cross terms (Yoma et al., 1998; Kitaoka and Nakagawa, 2002; Evans et al., 2006) in spectral subtraction. These studies, however, focused on improving speech recognition performance rather than improving speech quality. The study in Evans et al. (2006) assessed the effect of neglecting the cross terms on speech recognition performance. Significant degradations in performance were noted at SNR levels near 0 dB, but not at high SNR levels (>10 dB).

In the present paper, we take a new approach to spectral subtraction based on geometric principles. The proposed algorithm is based on a geometric approach (GA), and we will henceforth refer it to as the GA algorithm. It addresses the two aforementioned major shortcomings of spectral subtraction: the musical noise and invalid assumptions about the cross terms being zero. The approach taken is largely deterministic and is based on representing the noisy speech spectrum in the complex plane as the sum of the clean signal and noise vectors. Representing the noisy spectrum geometrically in the complex plane can provide valuable insights to the spectral subtraction approach that might otherwise not be obvious. For one, such geometric viewpoint can provide upper bounds on the difference between the phases of the noisy and clean spectra (Vary, 1985). It will also tell us whether it is theoretically possible to recover *exactly* the clean signal magnitude given the noisy speech spectrum, and under what conditions. Finally, it will inform us about the implications of discarding the cross terms in as far as obtaining accurate estimates of the magnitude spectrum.

This paper is organized as follows. Section 2 provides an overview of the power spectral subtraction algorithm and the assumptions made. Section 3 presents the proposed geometric algorithm, Section 4 provides the implementation details, Section 5 presents the simulation results and finally Section 6 presents our conclusions.

2. Power spectral subtraction: background and error analysis

Let $y(n) = x(n) + d(n)$ be the sampled noisy speech signal consisting of the clean signal $x(n)$ and the noise signal $d(n)$. Taking the short-time Fourier transform of $y(n)$, we get

$$Y(\omega_k) = X(\omega_k) + D(\omega_k) \tag{1}$$

for $\omega_k = 2\pi k/N$ and $k = 0,1,2,\ldots,N-1$, where N is the frame length in samples. To obtain the short-term power spectrum of the noisy speech, we multiply $Y(\omega_k)$ in the above equation by its conjugate $Y^*(\omega_k)$. In doing so, Eq. (1) becomes

$$\begin{aligned}|Y(\omega_k)|^2 &= |X(\omega_k)|^2 + |D(\omega_k)|^2 + X(\omega_k)\cdot D^*(\omega_k) \\ &\quad + X^*(\omega_k)D(\omega_k) \\ &= |X(\omega_k)|^2 + |D(\omega_k)|^2 + 2|X(\omega_k)| \\ &\quad \cdot |D(\omega_k)|\cos(\theta_X(k) - \theta_D(k)).\end{aligned} \tag{2}$$

The terms $|D(\omega_k)|^2$, $X(\omega_k)\cdot D^*(\omega_k)$ and $X^*(\omega_k)\cdot D(\omega_k)$ cannot be obtained directly and are approximated as $E\{|D(\omega_k)|^2\}$, $E\{X^*(\omega_k)\cdot D(\omega_k)\}$ and $E\{X(\omega_k)\cdot D^*(\omega_k)\}$, where $E[\cdot]$ denotes the expectation operator. Typically, $E\{|D(\omega_k)|^2\}$ is estimated during non-speech activity, and is denoted by $|\widehat{D}(\omega_k)|^2$. If we assume that $d(n)$ is zero mean and uncorrelated with the clean signal $x(n)$, then the terms $E\{X^*(\omega_k)\cdot D(\omega_k)\}$ and $E\{X(\omega_k)\cdot D^*(\omega_k)\}$ reduce to zero. Thus, from the above assumptions, the estimate of the clean speech power spectrum, denoted as $|\widehat{X}(\omega_k)|^2$, can be obtained by

$$|\widehat{X}(\omega_k)|^2 = |Y(\omega_k)|^2 - |\widehat{D}(\omega_k)|^2. \tag{3}$$

The above equation describes the so called *power spectrum subtraction* algorithm. The estimated power spectrum $|\widehat{X}(\omega_k)|^2$ in Eq. (3) is not guaranteed to be positive, but can be half-wave rectified. The enhanced signal is finally obtained by computing the inverse Fourier transform of $|\widehat{X}(\omega_k)|$ using the phase of the noisy speech signal.

Eq. (3) can also be written in the following form:

$$|\widehat{X}(\omega_k)|^2 = H^2(\omega_k)|Y(\omega_k)|^2, \tag{4}$$

where

$$H(\omega_k) = \sqrt{1 - \frac{|\widehat{D}(\omega_k)|^2}{|Y(\omega_k)|^2}} = \sqrt{\frac{\gamma(k)-1}{\gamma(k)}} \tag{5}$$

is the *gain* (or *suppression*) *function* and $\gamma(k) \triangleq |Y(\omega_k)|^2/|\widehat{D}(\omega_k)|^2$. Assuming that the cross terms in (2) are zero, $H(\omega_k)$ is always positive taking values in the range of $0 \leqslant H(\omega_k) \leqslant 1$.

The cross terms, however, are not necessarily zero and can in some instances be extremely large relative to $|Y(\omega_k)|^2$. To assess the error introduced by Eq. (3) when the cross terms are left out, we rewrite Eq. (2) as follows:

$$|Y(\omega_k)|^2 = |X(\omega_k)|^2 + |D(\omega_k)|^2 + \Delta Y(\omega_k) = |\widehat{Y}(\omega_k)|^2 + \Delta Y(\omega_k), \quad (6)$$

where $|\widehat{Y}(\omega_k)|^2 \triangleq |X(\omega_k)|^2 + |D(\omega_k)|^2$ and $\Delta Y(\omega_k)$ denotes the cross terms. From the above equation, we can define the following relative error introduced when neglecting the cross terms:

$$\varepsilon(k) \triangleq \frac{\left||Y(\omega_k)|^2 - |\widehat{Y}(\omega_k)|^2\right|}{|Y(\omega_k)|^2} = \frac{|\Delta Y(\omega_k)|}{|Y(\omega_k)|^2}. \quad (7)$$

Note that the above cross term error $\varepsilon(k)$ is normalized with respect to the noisy power spectrum of speech. Also, it is assumed in Eq. (7), that the noise spectrum $|D(\omega_k)|^2$ is known exactly. Consequently, the cross term error $\varepsilon(k)$ underestimates the true error incurred when in practice the noise spectrum needs to be estimated via a voice activity detection algorithm or a noise estimation algorithm.

Of great interest is finding out how this relative error $\varepsilon(k)$ varies as a function of the SNR at frequency bin k. The answer to this question will tell us about the range of SNR values for which the assumption that the cross terms are zero are valid. That is, it will tell us the conditions under which the power spectral subtraction rule Eq. (3) is accurate.

It is easy to show that the normalized cross term error $\varepsilon(k)$ can be written in terms of the SNR (in bin k) as follows (see proof in Appendix A):

$$\varepsilon(k) = \left|\frac{2\sqrt{\xi(k)}\cos(\theta_X(k) - \theta_D(k))}{1 + \xi(k) + 2\sqrt{\xi(k)}\cos(\theta_X(k) - \theta_D(k))}\right|, \quad (8)$$

where $\xi(k) \triangleq |X(\omega_k)|^2/|D(\omega_k)|^2$ denotes the true SNR in bin k. As expected, $\varepsilon(k) = 0$ when $\cos(\theta_X(k) - \theta_D(k)) = 0$, consistent with Eq. (2). From Eq. (8), we can see that $\varepsilon(k) \to 0$ when $\xi(k) \to \infty$ *or* when $\xi(k) \to 0$. Hence, asymptotically as the SNR $\to \pm\infty$, it is safe to make the assumption that the cross terms are negligible. For SNR values in between, however, the cross term error $\varepsilon(k)$ is not negligible and it can be quite large reaching a maximum when $\xi(k) = 1$ (i.e., SNR = 0 dB).

Fig. 1 plots $\varepsilon(k)$ as a function of $\xi(k)$ (expressed in dB) for fixed values of $\cos(\theta_X(k) - \theta_D(k))$. It is clear from this figure that $\varepsilon(k)$ is large for a wide range of $\xi(k)$ values centered around 0 dB, and particularly within the $[-20, 20]$ dB range. Outside this range, $\varepsilon(k)$ is extremely small. The error $\varepsilon(k)$ depends on the value of $\cos(\theta_X(k) - \theta_D(k))$, but is more sensitive to the values of $\cos(\theta_X(k) - \theta_D(k))$ for SNR values near 0 dB. The error is largest when $\cos(\theta_X(k) - \theta_D(k)) < 0$ and $\xi(k) \approx 1$ (i.e., SNR ≈ 0 dB), with $\varepsilon(k) = \infty$ when $\cos(\theta_X(k) - \theta_D(k)) = -1$ and $\xi(k) = 1$. The error is considerably smaller when $\cos(\theta_X(k) - \theta_D(k)) > 0$. In fact, it is bounded in the range $0 \leqslant \varepsilon(k) \leqslant 0.5$, with the upper bound attained when $\cos(\theta_X(k) - \theta_D(k)) = 1$ and $\xi(k) = 1$.

As it is evident from Fig. 1, the cross term error $\varepsilon(k)$ is large particularly when the SNR is near 0 dB. Unfortunately, this is the spectral SNR that most speech enhancement algorithms operate. To illustrate this, we show in Figs. 2 and 3 histograms (normalized) of $\varepsilon(k)$ obtained using real speech data embedded at low (5 dB SNR) and high (15 dB SNR) global SNR levels (estimated using the rms levels of speech and noise) in multi-talker babble and car noise respectively. These figures also show the corresponding histograms of $\xi(k)$ (in dB) for the same data and SNR levels (note that the histograms show $\xi(k)$ for all frequency bins). A total of 30 sentences (>1 min of speech) taken from the NOIZEUS corpus (Hu and Loizou, 2007) was used for the computation of these histograms. As shown in Figs. 2 and 3, the instantaneous SNR $\xi(k)$ has a wide distribution which spans the range of -40 dB to $+40$ dB, with a peak near 0 dB in both types of noise and SNR levels. Hence, the $\xi(k) = 0$ dB value is quite common even at high SNR levels. Examining the distribution of the cross term error $\varepsilon(k)$ (right column in Figs. 2 and 3), we note that it spans for the most part the range of [0, 0.5], with a small portion exceeding 0.5. As mentioned earlier (see also Eq. (7)), the error $\varepsilon(k)$ is expressed relative to the value of $|Y(\omega_k)|^2$ and is not absolute. So, if for instance, $\varepsilon(k) = 0.5$, then the magnitude of the cross terms will be 50% of the value of $|Y(\omega_k)|^2$, i.e., it will be quite significant. The fact that the distribution of $\varepsilon(k)$ is not concentrated at zero (see right panels in Fig. 2) provides further support to our hypothesis that the cross terms in Eq. (2) are not necessarily zero and should not be ignored.

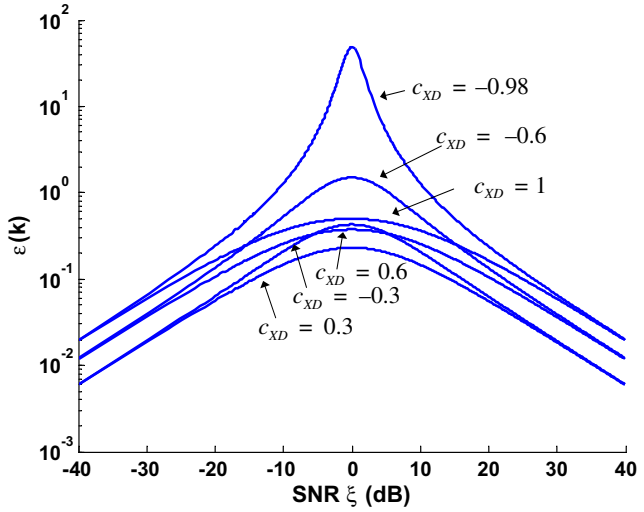

Fig. 1. Plot of the normalized cross-error term ε(k) as a function of spectral SNR, ξ, in dB for different values of c_{XD}, where $c_{XD} \triangleq \cos(\theta_X - \theta_D)$.

To summarize, the above error analysis suggests that the implicit assumption in Eq. (3) that the cross terms are zero is not valid for spectral SNR (i.e., $\xi(k)$) values near 0 dB, which is the region wherein most speech enhancement algorithms operate. Consequently, large estimation errors can result from the approximation given by Eq. (3). The conclusion that the cross term error $\varepsilon(k)$ is largest for SNR levels near 0 dB is consistent with the analysis in Evans et al.

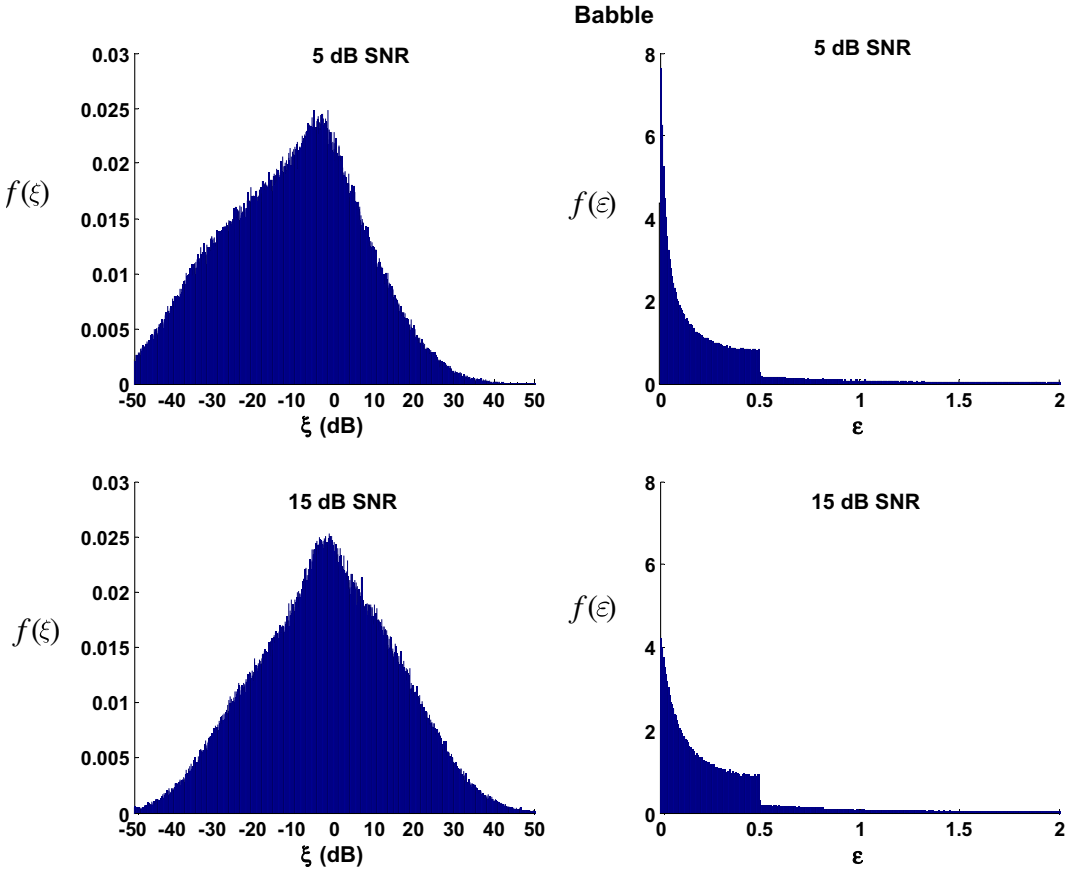

Fig. 2. Plots on the left show histograms of ξ (in dB) for speech embedded in multi-talker babble at 5 and 15 dB SNR. Plots on the right show histograms of the normalized cross-error term $\varepsilon(k)$ for speech embedded in multi-talker babble at 5 and 15 dB SNR.

(2006). Significant degradations in speech recognition performance were noted in Evans et al. (2006) for SNR levels near 0 dB, but not at high SNR levels (>10 dB). Next, we present a new algorithm that makes no assumptions about the cross terms in Eq. (2) being zero.

3. A geometric approach to spectral subtraction

From Eq. (1) we note that the noisy spectrum $Y(\omega_k)$ at frequency ω_k is obtained by summing two complex-valued spectra at frequency ω_k. As such, $Y(\omega_k)$ can be represented geometrically in the complex plane as the sum of two complex numbers, $X(\omega_k)$ and $D(\omega_k)$. This is illustrated in Fig. 4 which shows the representation of $Y(\omega_k)$ as a vector addition of $X(\omega_k)$ and $D(\omega_k)$ in the complex plane.

Eq. (5) gave the commonly used gain function of power spectrum subtraction algorithms that is obtained after making the assumption that the cross terms are zero or equivalently that the phase difference $[\theta_X(k) - \theta_D(k)]$ is equal to $\pm\pi/2$. Next, we derive the general gain function for spectral subtraction that makes no assumptions about the value of the phase difference between the noise and clean signals. We first rewrite Eq. (1) in polar form as

$$a_Y e^{j\theta_Y} = a_X e^{j\theta_X} + a_D e^{j\theta_D}, \tag{9}$$

where $\{a_Y, a_X, a_D\}$ are the magnitudes and $\{\theta_Y, \theta_X, \theta_D\}$ are the phases of the noisy, clean and noise spectra respectively. We henceforth drop the frequency index k for convenience.

Next, consider the triangle shown in Fig. 5. Using the Law of Sines or equivalently the right triangle $\overset{\triangle}{ABC}$ with $\overline{AB} \perp \overline{BC}$, we have

$$\begin{aligned}
\overline{AB} &= a_Y \sin(\theta_D - \theta_Y) = a_X \sin(\theta_D - \theta_X), \\
&\Rightarrow a_Y^2 \sin^2(\theta_D - \theta_Y) = a_X^2 \sin^2(\theta_D - \theta_X), \\
&\Rightarrow a_Y^2 [1 - \cos^2(\theta_D - \theta_Y)] = a_X^2 [1 - \cos^2(\theta_D - \theta_X)], \\
&\Rightarrow a_Y^2 (1 - c_{YD}^2) = a_X^2 (1 - c_{XD}^2),
\end{aligned} \tag{10}$$

where $c_{YD} \triangleq \cos(\theta_Y - \theta_D)$ and $c_{XD} \triangleq \cos(\theta_X - \theta_D)$. From the above equation, we can obtain the new gain function

$$H_{\mathrm{GA}} = \frac{a_X}{a_Y} = \sqrt{\frac{1 - c_{YD}^2}{1 - c_{XD}^2}}. \tag{11}$$

The above gain function is always real and positive (i.e., $H_{\mathrm{GA}} \geqslant 0$) since the terms c_{YD} and c_{XD} are bounded by one. Unlike the power spectral subtraction gain function Eq. (5) which is always positive and smaller (or equal) than one, the above gain function can be larger than one if

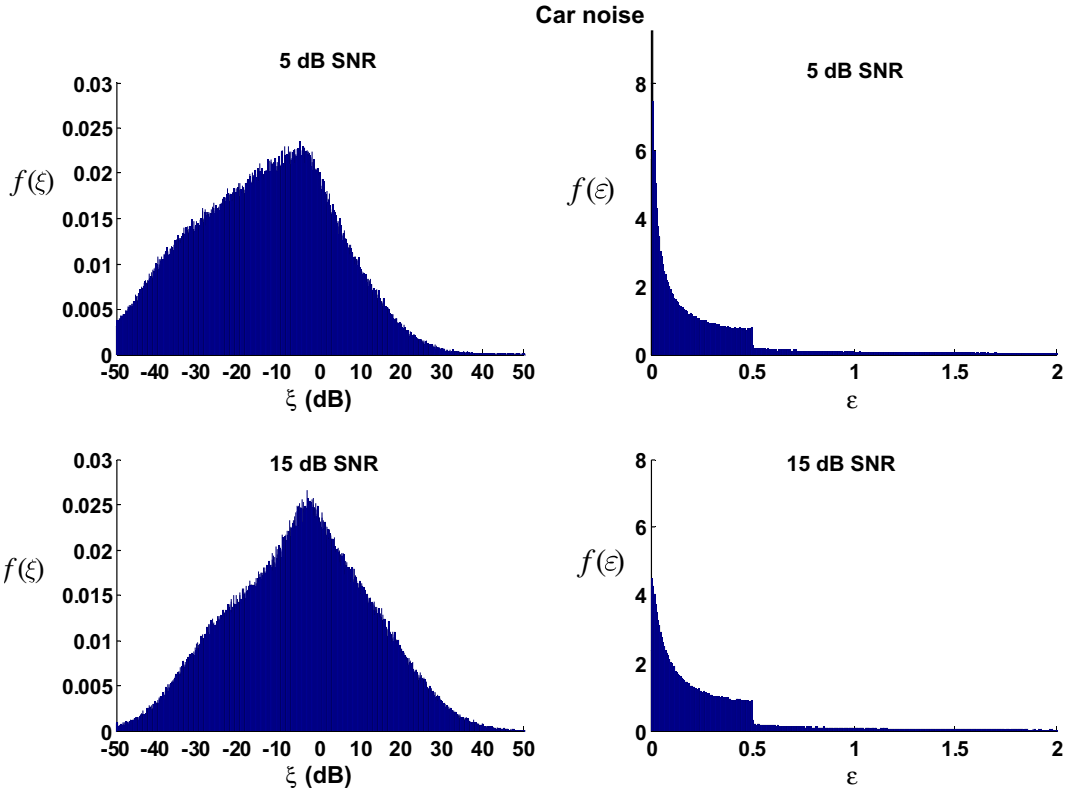

Fig. 3. Plots on the left are histograms of ξ (in dB) for speech embedded in car noise at 5 and 15 dB SNR. Plots on the right are histograms of the normalized cross-error term $\varepsilon(k)$ for speech embedded in car noise at 5 and 15 dB SNR.

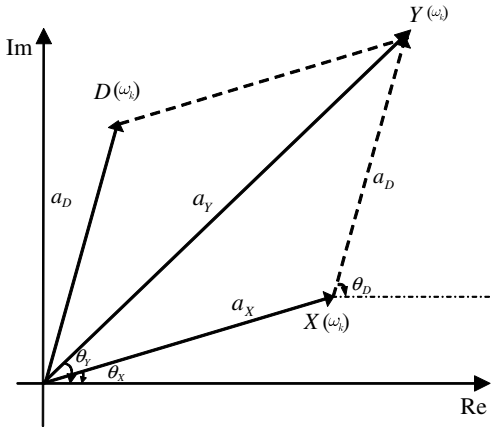

Fig. 4. Representation of the noisy spectrum $Y(\omega_k)$ in the complex plane as the sum of the clean signal spectrum $X(\omega_k)$ and noise spectrum $D(\omega_k)$.

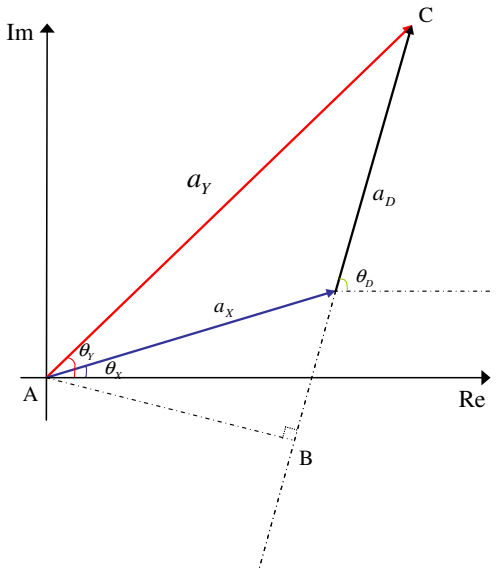

Fig. 5. Triangle showing the geometric relationship between the phases of the noisy speech, noise and clean speech spectra.

$|c_{YD}| < |c_{XD}|$. Eq. (11) is one of many equations that can be derived using trigonometric principles. Alternative equations can be found in Loizou (2007, Ch. 5).

It is worthwhile noting that the above suppression function reduces to the suppression function of the power spectral subtraction method (i.e., Eq. (5)) if $c_{XD} = 0$, i.e., if the signal and noise vectors are orthogonal to each other. Statistically, if the signal and noise are orthogonal to each other (i.e., $E[X(\omega_k) \cdot D(\omega_k)] = 0$) and are zero mean, then they are also uncorrelated (Papoulis and Pillai, 2002, p. 211). To prove that the above suppression function reduces to that given in Eq. (5) when $c_{XD} = 0$, it is easy to see from Fig. 4 that when the noise and clean signal vectors are orthogonal to each other (i.e., $c_{XD} = 0$), then

$$c_{YD} = \frac{a_D}{a_Y}. \tag{12}$$

Substituting the above equation in Eq. (11), we get Eq. (5). In view of the above analysis, we can say that the suppression rule given in Eq. (11) is the true and exact suppression rule for spectral subtractive algorithms if no assumptions are made about the statistical relationship between the signal and the noise. In contrast, the suppression rule given in Eq. (5) is merely an approximation since it assumes that $c_{XD} = 0$, i.e., that the clean signal and noise vectors are orthogonal to each other over short-time intervals (20–30 ms). Multiplication of the noisy signal by the suppression function given in Eq. (5) would not yield the clean signal magnitude spectrum even if we had access to the true noise magnitude spectrum (i.e., $|D(\omega_k)|$). In contrast, multiplication of the noisy magnitude spectrum (a_Y) by the suppression function given in Eq. (11) would yield exactly the clean signal magnitude spectrum (i.e., a_X).

The aforementioned suppression function relies on the estimation of the phase differences between the noisy (or clean) and noise signals. That by itself, however, is a difficult task and no methods currently exist to determine the values of these phases accurately. One possibility is to derive and make use of explicit relationships between the phases of noisy and noise signals using trigonometric principles. In doing so, we can solve explicitly for c_{YD} and c_{XD} yielding (see proof in Appendix B)

$$c_{YD} = \frac{a_Y^2 + a_D^2 - a_X^2}{2a_Y a_D}, \tag{13}$$

$$c_{XD} = \frac{a_Y^2 - a_X^2 - a_D^2}{2a_X a_D}. \tag{14}$$

Clearly, the main obstacle in utilizing the above equations to estimate the phase differences between the signal and noise signals is their dependency on the clean signal amplitude, which we do not have. We can however derive an alternative equation for c_{YD} and c_{XD} by dividing both numerator and denominator of Eqs. (13) and (14) by a_D^2. In doing so, we get

$$c_{YD} = \frac{\gamma + 1 - \xi}{2\sqrt{\gamma}}, \tag{15}$$

$$c_{XD} = \frac{\gamma - 1 - \xi}{2\sqrt{\xi}}, \tag{16}$$

where the variables γ and ξ are defined as follows:

$$\gamma \triangleq \frac{a_Y^2}{a_D^2}, \tag{17}$$

$$\xi \triangleq \frac{a_X^2}{a_D^2}. \tag{18}$$

Note that the terms γ and ξ are the instantaneous versions of the *a posteriori* and *a priori* SNRs, respectively used in MMSE algorithms (Ephraim and Malah, 1984; Loizou, 2005). Substituting Eqs. (15) and (16) into Eq. (11), we get the following expression for the suppression function in terms of γ and ξ:

$$H_{\mathrm{GA}}(\xi,\gamma) = \sqrt{\frac{1 - \frac{(\gamma+1-\xi)^2}{4\gamma}}{1 - \frac{(\gamma-1-\xi)^2}{4\xi}}}. \tag{19}$$

The above suppression function can in principle be larger than one. Much like the gain function of MMSE-based enhancement algorithms (Ephraim and Malah, 1984), the above gain function depends on two parameters, γ and ξ. To better understand the dependency of these two parameters on suppression, we plot in Fig. 6, $H_{\mathrm{GA}}(\xi,\gamma)$ as a function of $(\gamma - 1)$ for fixed values of ξ. The suppression curves of the MMSE algorithm (Ephraim and Malah, 1984) are superimposed for comparison. It is clear that the gain functions of the GA algorithm follow for the most part the pattern of the MMSE gain functions. For values of $(\gamma - 1)$ smaller than 5 dB, the gain functions of the GA approach follow closely the MMSE gain functions, and deviate thereafter. For values of $(\gamma - 1)$ greater than 5 dB, the gain functions of the GA algorithm become increasing more suppressive than the MMSE gain functions.

Fig. 7 plots $H_{\mathrm{GA}}(\xi,\gamma)$ as a function of ξ for fixed values of $(\gamma - 1)$. The Wiener gain function (e.g., $H_W = \xi/(\xi + 1)$) and the MMSE gain function are also plotted for comparison. Overall, for the same value of $(\gamma - 1)$, the GA gain function follows closely the MMSE gain function for small and negative values of $(\gamma - 1)$. The GA gain function, however, becomes more suppressive than the MMSE gain function for values of $(\gamma - 1)$ larger than 5 dB, consistent with the suppression curves in Fig. 6.

The fact that the gain function of the GA algorithm has similar characteristics as those found in the MMSE algorithms, suggests that it inherits the properties and behavior of the MMSE algorithm (Cappe, 1994). Much like in the MMSE algorithm, the *a posteriori* parameter γ acts as a correction parameter that influences attenuation only when ξ is low. The correction, however, is done in an intuitively opposite direction. As shown in Fig. 6, strong attenuation is applied when γ is large, and not when γ is small as one would expect. This counter-intuitive behavior is not an artifact of the algorithm, but it is in fact useful when dealing with low-energy speech segments. In segments containing background noise, the γ values are in some frames unrealistically high, and those frames are assigned an increased attenuation. This over-attenuation is done because the suppression rule puts more "faith" in the ξ values, which are small in those frames compared to the γ values. Since the attenuation in the MMSE algorithm is primarily influenced by the smoothed value of the *a priori* SNR, the attenuation itself will not change radically from frame to frame. Consequently, the musical noise will be reduced or eliminated altogether. In contrast, the standard power spectral subtraction algorithm depends on the estimation of the *a posteriori* SNR which can change radically from frame to frame. As a result, musical noise is produced. In summary, it is the smoothing behavior of the "decision-directed" approach in conjunction with the MMSE suppression rule that is responsible for reducing

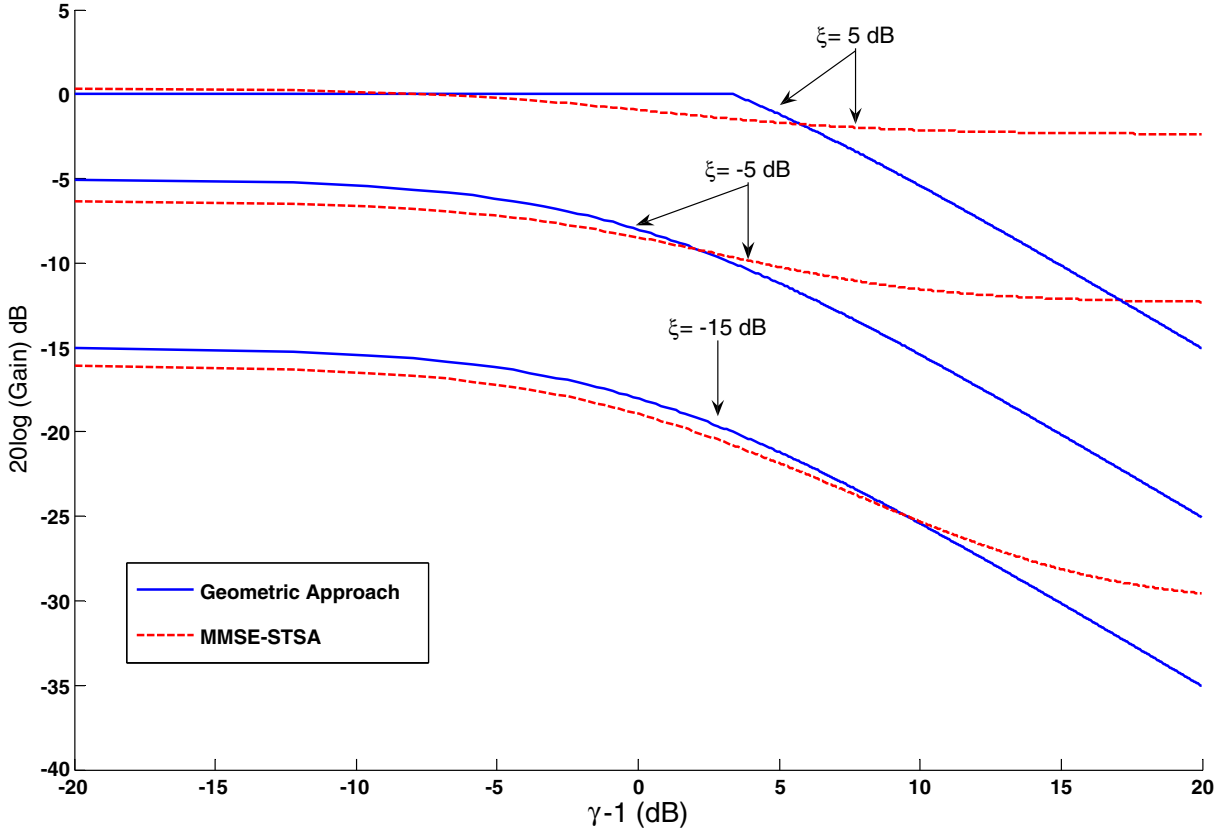

Fig. 6. Plot of the suppression curves (solid lines) of $H_{GA}(\xi, \gamma)$ as a function of $(\gamma - 1)$ for fixed values of ξ. The corresponding suppression curves (dashed lines) of the MMSE algorithm (Ephraim and Malah, 1984) are superimposed for comparison.

the musical noise effect in the MMSE algorithm (Cappe, 1994). Since the GA algorithm inherits the behavior of the MMSE algorithm, we expect little or no musical noise with the GA algorithm.

We should point out here that there are two main differences between the proposed GA algorithm and the MMSE algorithm. First, the GA algorithm is deterministic and is not derived using any statistical model. The clean magnitude spectrum is treated as unknown, but deterministic. Consequently, no assumptions are made about the statistical distributions of the speech and noise Fourier transform coefficients, as done in the MMSE algorithm. Second, the parameters γ and ξ are instantaneous values and not long-term, statistical average values. Consequently, different techniques need to be employed to estimate these parameters, and these techniques are discussed next. For completeness, we assess and compare the performance of the proposed GA algorithm using both instantaneous and long-term average measurements of γ and ξ.

4. Implementation

The gain function given in Eq. (19) is the ideal one and in practice it needs to be estimated from the noisy observations. The implementation of the gain function requires estimates of γ and ξ. According to Eqs. (17) and (18), γ and ξ are instantaneous values and not long-term, statistical average values as in Ephraim and Malah (1984). Note that in Ephraim and Malah (1984), these two terms were defined as $\xi_M \triangleq E[a_X^2]/E[a_D^2]$ and $\gamma_M \triangleq a_Y^2/E[a_D^2]$. Therefore, the methods proposed in Ephraim and Malah (1984) cannot be used to estimate γ and ξ in Eq. (19). Alternative methods are thus proposed in this paper for estimating γ and ξ.

To estimate ξ, we propose to use present as well as past spectral information. More specifically, we can make use of the enhanced magnitude spectrum obtained in the past frame and approximate ξ as

$$\hat{\xi}_I(\lambda, k) = \hat{a}_X^2(\lambda - 1, k)/\hat{a}_D^2(\lambda - 1, k), \tag{20}$$

where $\hat{\xi}_I(\lambda, k)$ indicates the estimate of ξ at frame λ and bin k, and the subscript I indicates instantaneous measurement. The above estimate of the instantaneous value of ξ utilizes only (immediate) past spectral information. We can also utilize the relationship between the true values of γ and ξ (see Appendix B, Eq. (29)) and get an estimate of ξ based on spectral information available in the present frame. More precisely, as shown in Appendix B, $\xi = \gamma + 1 - 2\sqrt{\gamma} \cdot c_{YD}$, and after exploiting the fact that c_{YD} is bounded, we can use the lower bound of ξ (see Appendix B, Eq. (29)) as its estimate, i.e., $\hat{\xi}(\lambda, k) = \left(\sqrt{\hat{\gamma}(\lambda, k)} - 1\right)^2$, where $\hat{\gamma}(\lambda, k)$ denotes the estimate of γ at frame λ and bin k. Combining the two estimates of ξ derived using past and present spectral information, we get

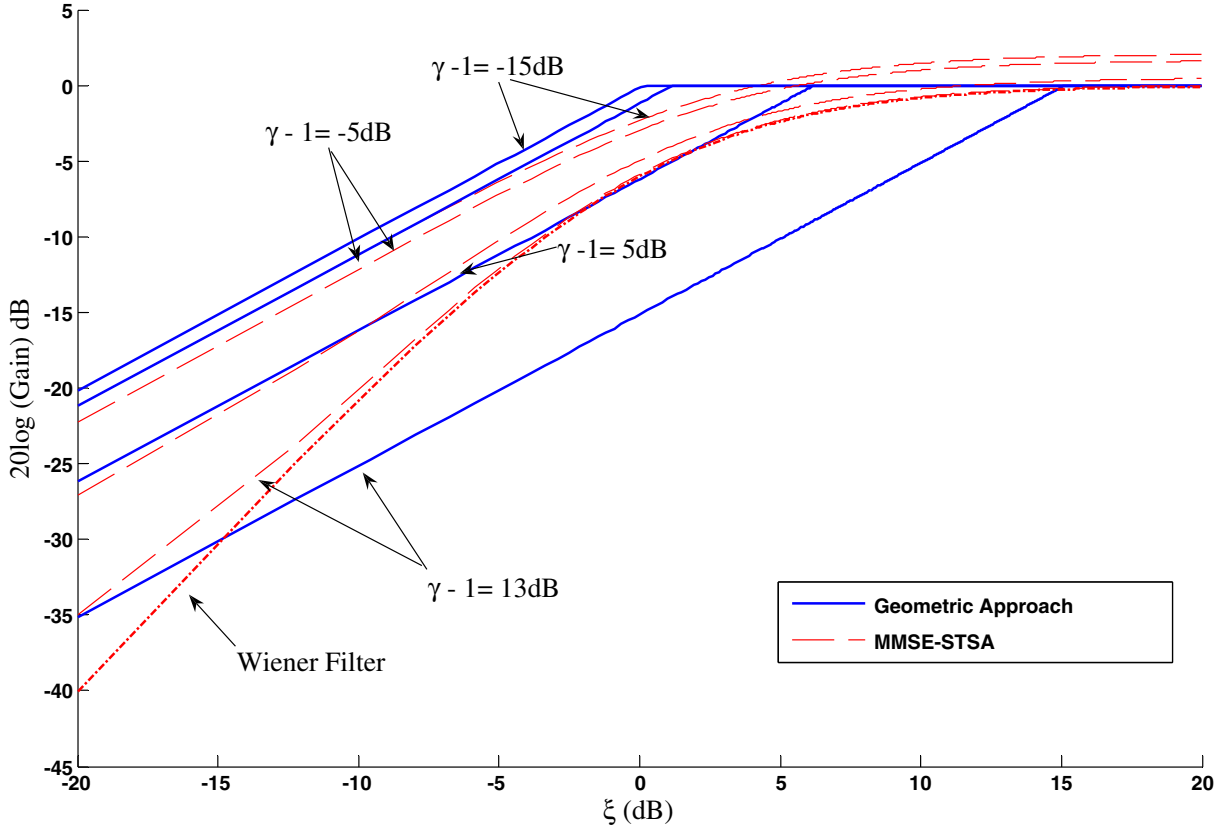

Fig. 7. Plots of the suppression curves of $H_{GA}(\xi,\gamma)$ as a function of ξ for fixed values of $(\gamma-1)$. The corresponding suppression curves (dashed lines) of the MMSE algorithm (Ephraim and Malah, 1984) and Wiener algorithm are superimposed for comparison

$$\hat{\xi}(\lambda,k)=\alpha\cdot\left[\frac{\hat{a}_X(\lambda-1,k)}{\hat{a}_D(\lambda-1,k)}\right]^2+(1-\alpha)\cdot\left(\sqrt{\hat{\gamma}(\lambda,k)}-1\right)^2, \tag{21}$$

where α is a smoothing constant, and $\hat{a}_D(\lambda,k)$ is the estimate of the magnitude spectrum of the noise. The above equation is a weighted average of past and present SNR instantaneous measurements, and the smoothing constant controls the weight placed on past and present spectral information. Note that Eq. (21) gives an average estimate of ξ utilizing past and present spectral information. On that regard it is similar to the decision-directed approach used in Ephraim and Malah (1984). If $\alpha=1$, then Eq. (21) reduces to the instantaneous estimate of ξ given in Eq. (20). Both estimates (instantaneous and average) of ξ will be explored and evaluated.

For $\hat{\gamma}(\lambda,k)$, we use the following instantaneous estimate:

$$\hat{\gamma}_I(\lambda,k)=\left(\frac{a_Y(\lambda,k)}{\hat{a}_D(\lambda,k)}\right)^2, \tag{22}$$

where $\hat{a}_D(\lambda,k)$ is an estimate of the noise spectrum obtained using a noise-estimation algorithm. We considered smoothing and limiting the values of $\hat{\gamma}(\lambda,k)$ in order to reduce the rapid fluctuations associated with the above computation of $\hat{\gamma}(\lambda,k)$ and also to limit the over-suppression of the signal for large values of $\hat{\gamma}(\lambda,k)$ (see Fig. 6). We consider smoothing $\hat{\gamma}(\lambda,k)$ as follows:

$$\hat{\gamma}_{GA}(\lambda,k)=\beta\cdot\hat{\gamma}_{GA}(\lambda-1,k)+(1-\beta)\cdot\min[\hat{\gamma}_I(\lambda,k),20], \tag{23}$$

where $\hat{\gamma}_{GA}(\lambda,k)$ is the smoothed estimate of γ, $\hat{\gamma}_I(\lambda,k)$ is given by Eq. (22) and β is a smoothing constant. The *min* operation was used to limit the value of $\hat{\gamma}_I(\lambda,k)$ to a maximum of 13 dB ($=10\log_{10}(20)$) and avoid over-attenuation of the signal (see Fig. 6). Note that when $\beta=0$ in Eq. (23), we get $\hat{\gamma}_{GA}(\lambda,k)=\hat{\gamma}_I(\lambda,k)$. We found that the smoothing of $\hat{\gamma}(\lambda,k)$ improved the estimate of $\hat{a}_X(\lambda,k)$ in the mean-square error sense (see experiments in the next section). Note also that Eq. (21) gives an average estimate of γ utilizing past and present spectral information. If $\beta=0$, then Eq. (23) reduces to the instantaneous estimate of γ given in Eq. (22). Both estimates (instantaneous and average) of γ will be explored and evaluated.

The above estimated values of γ and ξ (i.e., $\hat{\gamma}_{GA}(\lambda,k)$ and $\hat{\xi}(\lambda,k)$) are used to approximate the gain function in Eq. (19). In principle, the transfer function given in Eq. (19) is based on instantaneous values of γ and ξ. In practice, however, the true values of γ and ξ may vary drastically from frame to frame and it is extremely challenging to estimate those values with high degree of accuracy and reliability. Furthermore, we cannot compute the true value of ξ as we lack access to the clean signal spectrum. We are thus forced to rely on past estimates of the clean signal spectrum to approximate ξ. Given that γ and ξ can be estimated

using either instantaneous (e.g., Eq. (22)) or average estimates (e.g., Eq. (23)), we will explore both possibilities. In doing so, we will make use of two transfer functions. The first transfer function, denoted as $\widehat{H}_{\mathrm{GA}_I}(\hat{\xi}_I, \hat{\gamma}_I)$, is based on the instantaneous measurements of γ and ξ given by Eq. (22) and Eq. (20) respectively. The second transfer function, denoted as $\widehat{H}_{\mathrm{GA}}(\hat{\xi}, \hat{\gamma}_{\mathrm{GA}})$, is based on the long-term average measurements of γ and ξ given by Eqs (23), (21) respectively. Both transfer functions will be explored and compared.

As mentioned earlier, $H_{\mathrm{GA}}(\xi,\gamma)$ (as well as $\widehat{H}_{\mathrm{GA}}(\hat{\xi}, \hat{\gamma}_{\mathrm{GA}})$ or $\widehat{H}_{\mathrm{GA}_I}(\hat{\xi}_I, \hat{\gamma}_I)$) can be larger than one. From Eq. (2) we see that $H_{\mathrm{GA}}(\xi,\gamma) > 1$ when $c_{XD} < 0$. But, as shown in Fig. 1, when $c_{XD} < 0$, the normalized cross term error $\varepsilon(k)$ can be large, particularly if $\xi \approx 0$ dB. This suggests that the cross term error $\varepsilon(k)$, and possibly the magnitude spectrum estimation error, can be large when $\widehat{H}_{\mathrm{GA}}(\hat{\xi}, \hat{\gamma}_{\mathrm{GA}}) > 1$. For that reason, we decided to limit the value of $\widehat{H}_{\mathrm{GA}}(\hat{\xi}, \hat{\gamma}_{\mathrm{GA}})$ to be always smaller (or equal) to 1.

To summarize, the proposed GA algorithm involved the following steps, which were applied to each frame of noisy speech

Step 1: Using the FFT, compute the magnitude spectrum $a_Y(\lambda,k)$ of the noisy signal at frame λ.
Step 2: Using a noise estimation algorithm (e.g., Martin, 2001), update the power spectrum of the noise signal, i.e., update $[\hat{a}_D(\lambda,k)]^2$.
Step 3: Compute $\hat{\gamma}_{\mathrm{GA}}(\lambda,k)$ according to Eqs. (23) and (22).
Step 4: Use $\hat{\gamma}_{\mathrm{GA}}(\lambda,k)$ to estimate $\hat{\xi}(\lambda,k)$ according to Eq. (21). Floor $\hat{\xi}(\lambda,k)$ to $\xi_{\min}$ for values of $\hat{\xi}(\lambda,k)$ smaller than $\xi_{\min}$, where $\xi_{\min} = -26$ dB.
Step 5: Estimate the gain function $\widehat{H}_{\mathrm{GA}}(\hat{\xi}, \hat{\gamma}_{\mathrm{GA}})$ using Eq. (19) and limit it to 1.
Step 6: Obtain the enhanced magnitude spectrum of the signal by: $\hat{a}_X(\lambda,k) = \widehat{H}_{\mathrm{GA}}(\hat{\xi}, \hat{\gamma}_{\mathrm{GA}}) \cdot a_Y(\lambda,k)$.
Step 7: Compute the inverse FFT of $\hat{a}_X(\lambda,k) \cdot e^{j\theta_Y(\lambda,k)}$, where $\theta_Y(\lambda,k)$ is the phase of the noisy signal, to obtain the enhanced speech signal.

The above algorithm uses the transfer function $\widehat{H}_{\mathrm{GA}}(\hat{\xi}, \hat{\gamma}_{\mathrm{GA}})$ that is based on smoothed measurements of γ and ξ (Eqs. (21) and (23)). The algorithm which uses the transfer function $\widehat{H}_{\mathrm{GA}_I}(\hat{\xi}_I, \hat{\gamma}_I)$, based on instantaneous measurements of γ and ξ, can be implemented in a similar fashion by setting $\beta = 0$ in Eq. (23) and $\alpha = 1$ in Eq. (21). We will be referring to the instantaneous version of the GA algorithm as the GAi algorithm.

The proposed GA algorithm was applied to 20-ms duration frames of speech using a Hamming window, with 50% overlap between frames. The overlap-and-add method was used to reconstruct the enhanced signal. The smoothing constants used in Eqs. (21) and (23) were set to $\alpha = 0.98$ and $\beta = 0.6$, respectively. These constants were chosen based on listening experiments as well as experiments assessing the mean-square error between the estimated and true magnitude spectra (see evaluation in next section). For the GAi algorithm, these constants were set to $\alpha = 1$ and $\beta = 0$. The minimum statistics noise estimation algorithm (Martin, 2001) was used for estimating/updating the noise spectrum.

5. Evaluation

We performed two types of evaluation of the proposed GA algorithm. In the first evaluation, we computed the mean square error (MSE) between the estimated (enhanced) magnitude ($\hat{a}_X$) and the true (clean) magnitude spectra (a_X). This evaluation was done for both the proposed algorithm and the traditional spectral-subtraction algorithm which assumes that the cross terms are zero. This comparison will tell us whether the MSE value will be reduced when the cross terms are taken into account. Small values of MSE, however, might not necessarily imply better speech quality as the MSE is not a perceptually motivated error criterion (Loizou, 2005). For that reason, we conduct a second evaluation in which we compare the quality of enhanced speech by the proposed and standard spectral subtractive algorithms using objective measures.

5.1. MSE evaluation

The MSE between the true and estimated magnitude spectra is defined as

$$\mathrm{MSE} = \frac{1}{M \cdot N} \sum_{\lambda=0}^{M-1} \sum_{k=0}^{N-1} (a_X(\lambda,k) - \hat{a}_X(\lambda,k))^2, \tag{24}$$

where $a_X(\lambda,k)$ is the true (clean) magnitude spectrum at frame λ and bin k, $\hat{a}_X(\lambda,k)$ is the estimated magnitude spectrum (following enhancement), M is the total number of frames in a sentence, and N is the number of frequency bins.

The MSE was computed for the proposed algorithm and compared against the corresponding MSE values obtained by the traditional spectral subtraction algorithm. To assess whether smoothing $\hat{\gamma}(\lambda,k)$ as per Eq. (23) provided any significant reductions in MSE, we also conducted another set of experiments in which we varied the smoothing constant β Eq. (23) from 0 to 1. A value of $\beta = 0$ corresponds to no smoothing of $\hat{\gamma}(\lambda,k)$. The smoothing constant α was fixed at $\alpha = 0.98$ (this was based on prior experiments demonstrating that best performance is obtained with high values (close to one) of α. As an additional comparison, we included the evaluation of the GAi algorithm in which $\beta = 0$ and $\alpha = 1$. In fairness, we also implemented the basic spectral subtractive algorithm given in Eq. (5), and replaced $\hat{\gamma}(k)$ in Eq. (5) with its smoothed version given in Eq. (23). We refer to this algorithm as SSsm. Thirty sentences from the NOIZEUS database (Hu and Loizou, 2007) were used for the MSE evaluation of the proposed algorithm. The NOIZEUS sentences[1] were sampled at

[1] Available from: http://www.utdallas.edu/~loizou/speech/noizeus/.

Table 1
MSE values obtained by the GA algorithm, the GAi algorithm ($\alpha = 1$, $\beta = 0$), the spectral subtraction (SS) algorithm and the smoothed spectral subtractive (SSsm) algorithm for different values of β

Algorithm	SNR (dB)	$\beta = 0$	$\beta = 0.2$	$\beta = 0.4$	$\beta = 0.6$	$\beta = 0.8$	$\beta = 0.98$	$\beta = 0$, $\alpha = 1$
GA	0	2.74	2.70	2.72	2.67	2.41	1.97	4.51
SSsm	0	4.26	4.22	4.11	3.95	3.79	3.76	4.26
SS	0	4.26	4.26	4.26	4.26	4.26	4.26	4.26
GA	5	1.49	1.53	1.55	1.52	1.35	1.07	2.78
SSsm	5	1.33	1.31	1.28	1.24	1.21	1.32	1.33
SS	5	1.33	1.33	1.33	1.33	1.33	1.33	1.33
GA	10	0.81	0.86	0.87	0.85	0.76	0.59	1.60
SSsm	10	0.41	0.41	0.40	0.39	0.39	0.49	0.41
SS	10	0.41	0.41	0.41	0.41	0.41	0.41	0.41

8 kHz, and were corrupted by white noise at 0, 5 and 10 dB SNR levels. The results are given in Table 1.

It is clear from Table 1 that the proposed GA algorithm produced significantly smaller MSE values than the traditional spectral subtraction algorithm, and the difference was particularly evident at low SNR levels (0 and 5 dB) with $\beta = 0.98$. The MSE values obtained by the two algorithms at 10 dB were comparable, with the traditional spectral subtraction algorithm yielding slightly smaller MSE values. Best performance (smaller MSE values) was obtained with the GA algorithm with $\beta = 0.98$, clearly indicating that the smoothing of $\hat{\gamma}(\lambda, k)$ helped reduce the MSE. The benefit brought by smoothing $\hat{\gamma}(\lambda, k)$ in the SS algorithm was relatively small. Worst performance (larger MSE values) were obtained by the GA algorithm when γ and ξ were not smoothed (i.e., when instantaneous measurements were used) suggesting that inclusion of past and present spectral information can be beneficial. The outcome that the SS algorithm performed reasonably well and its performance was comparable to that obtained by the GA algorithm is consistent with the earlier observation (see Fig. 1) that the cross terms are negligible and can be ignored when the SNR is high. In stark contrast, the cross terms cannot be ignored when the SNR is near 0 dB (but can be ignored for extremely low SNR, i.e., SNR $\rightarrow -\infty$). In brief, the derived gain function Eq. (19) remains robust even at low SNR levels (0–5 dB), whereas the SS gain function Eq. (5) becomes inaccurate at low SNR levels (0 dB).

5.2. Quality evaluation

The proposed geometric approach (GA) algorithm was evaluated using the PESQ and log likelihood ratio (LLR) objective measures, which were found in Hu and Loizou (2008) to correlate moderately high with subjective judgments of speech quality. Thirty sentences from the NOIZEUS database (Hu and Loizou, 2007) were used for the objective evaluation of the proposed algorithm, with half of the sentences produced by 3 female speakers and the other half produced by 3 male speakers. The NOIZEUS sentences were sampled at 8 kHz, and were corrupted by multi-talker babble, street, and car noise taken from the AURORA database (Hirsch and Pearce, 2000) at 0, 5 and 10 dB SNR levels. The sentences were also corrupted by white noise at 0, 5 and 10 dB SNR levels.

The PESQ (ITU, 2000) and LLR objective measures (Hu and Loizou, 2006) were used to assess speech quality. The PESQ measure obtained a correlation of $\rho = 0.67$ in predicting overall quality of noise-suppressed speech (Hu and Loizou, 2006; Hu and Loizou, 2008), and the LLR measure obtained a correlation of $\rho = 0.61$. Higher correlations were obtained with the PESQ ($\rho = 0.89$) and LLR ($\rho = 0.85$) measures in Hu and Loizou (2008) after averaging objective scores and ratings across the various noise conditions. The segmental SNR measure, which is often used to evaluate the performance of speech enhancement algorithms, performed very poorly ($\rho = 0.31$) in Hu and Loizou (2008) and was therefore not used in this study.

For comparative purposes, we evaluate the performance of the traditional spectral subtraction (SS) algorithm implemented using Eq. (5), and implemented using the smoothed version of $\hat{\gamma}(k)$ given in Eq. (23). We refer to the latter implementation as the SSsm algorithm. In fairness, we used the same smoothing constant β as in the GA algorithm. For completeness, and for reference purposes only[2], we report the performance of the traditional MMSE algorithm (Ephraim and Malah, 1984) along with an implementation based on a smoothed version of $\hat{\gamma}(k)$ (i.e., Eq. (23)), which we refer to as the MMSEsm algorithm. The decision-directed approach was used in the implementation of the MMSE algorithm to estimate ξ with $\alpha = 0.98$. All algorithms were tested using two different values of β ($\beta = 0.6$ and $\beta = 0.98$) and with $\alpha = 0.98$. The latter value of β was found (see Table 1) to yield smaller MSE values than the traditional spectral subtraction algorithm.

[2] The objective evaluation of the MMSE algorithm is only included in this paper for completeness; as it shares some of its properties with the GA algorithm (see Section 3). The MMSE algorithm cannot be directly (or fairly) compared with the GA algorithm as it is based on different principles, designed using different assumptions and belonging to a different class of algorithms, namely the statistical-model based algorithms.

Table 2
Objective evaluation (in terms of PESQ values) and comparison of the proposed GA algorithm against the spectral subtraction (SS), the smoothed spectral subtractive (SSsm) algorithm, the MMSE (Ephraim and Malah, 1984) algorithm and the smoothed MMSE algorithm (MMSEsm)

Algorithm	Noise type	SNR = 0 dB		SNR = 5 dB		SNR = 10 dB	
		$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$
GA	Babble	1.81	1.62	2.16	2.07	2.50	2.44
GAi		1.65		1.88		2.14	
SS		1.73	1.73	2.04	2.04	2.37	2.37
SSsm		1.75	1.72	2.06	2.03	2.39	2.34
MMSE		1.76	1.76	2.12	2.12	2.51	2.51
MMSEsm		1.73	1.80	2.05	2.12	2.38	2.43
GA	Car	1.84	1.81	2.19	2.18	2.51	2.52
GAi		1.69		1.74		2.04	
SS		1.69	1.69	1.98	1.98	2.31	2.31
SSsm		1.72	1.67	2.00	1.93	2.34	2.23
MMSE		1.93	1.93	2.28	2.28	2.66	2.66
MMSEsm		1.88	1.81	2.16	2.15	2.45	2.43
GA	Street	1.76	1.69	2.16	2.11	2.50	2.47
GAi		1.43		1.79		2.11	
SS		1.70	1.70	2.00	2.00	2.36	2.36
SSsm		1.71	1.66	2.02	1.94	2.37	2.27
MMSE		1.80	1.80	2.20	2.20	2.58	2.58
MMSEsm		1.75	1.76	2.10	2.08	2.38	2.41
GA	White	1.81	1.90	2.20	2.24	2.53	2.58
GAi		1.47		1.77		2.08	
SS		1.66	1.66	1.95	1.95	2.29	2.29
SSsm		1.69	1.61	1.98	1.88	2.31	2.19
MMSE		2.00	2.00	2.39	2.39	2.74	2.74
MMSEsm		1.85	1.74	2.17	2.10	2.45	2.42

Table 3
Objective evaluation (in terms of LLR values) and comparison of the proposed GA algorithm against the spectral subtraction (SS), the smoothed spectral subtractive (SSsm) algorithm, the MMSE (Ephraim and Malah, 1984) algorithm and the smoothed MMSE algorithm (MMSEsm)

Algorithm	Noise type	SNR = 0 dB		SNR = 5 dB		SNR = 10 dB	
		$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$	$\beta = 0.6$	$\beta = 0.98$
GA	Babble	1.06	1.13	0.86	0.91	0.69	0.70
GAi		1.19		1.02		0.81	
SS		0.94	0.94	0.75	0.75	0.55	0.55
SSsm		0.93	0.90	0.74	0.72	0.54	0.53
MMSE		1.15	1.15	0.90	0.90	0.67	0.67
MMSEsm		1.24	1.15	1.00	0.90	0.83	0.70
GA	Car	0.98	1.04	0.80	0.83	0.66	0.65
GAi		1.24		1.05		0.87	
SS		1.00	1.00	0.78	0.78	0.59	0.59
SSsm		0.98	0.98	0.76	0.77	0.57	0.58
MMSE		1.01	1.01	0.79	0.79	0.63	0.63
MMSEsm		1.12	1.10	0.92	0.87	0.78	0.70
GA	Street	1.04	1.12	0.84	0.88	0.70	0.71
GAi		1.23		1.05		0.85	
SS		1.01	1.01	0.81	0.81	0.62	0.62
SSsm		1.00	0.98	0.79	0.80	0.60	0.61
MMSE		1.12	1.12	0.88	0.88	0.68	0.68
MMSEsm		1.27	1.19	1.04	0.95	0.89	0.76
GA	White	1.55	1.60	1.28	1.31	1.09	1.09
GAi		1.73		1.43		1.20	
SS		1.74	1.74	1.47	1.47	1.22	1.22
SSsm		1.72	1.66	1.45	1.39	1.20	1.15
MMSE		1.54	1.54	1.25	1.25	1.05	1.05
MMSEsm		1.86	1.83	1.57	1.50	1.37	1.25

The objective results are given in Table 2 for the PESQ measure and in Table 3 for the LLR measure. High values of PESQ indicate better performance, while high values of LLR indicate poor performance. The GA algorithm, implemented with $\beta = 0.6$, performed significantly and consistently better than the spectral subtractive algorithms (SS and SSsm) in all conditions. Statistical analysis (paired samples t-tests) confirmed that the differences were statistically significant ($p < 0.05$). The GA algorithm performed relatively worse when implemented with $\beta = 0.98$, particularly at the low-SNR levels. This suggests that the GA algorithm is sensitive to the value of β used for estimating and updating $\hat{\gamma}(\lambda, k)$. A value of $\beta = 0.6$ provides roughly equal weight to the use of past and spectral information when estimating $\hat{\gamma}(\lambda, k)$. In contrast, the performance of the spectral subtractive algorithm was not affected significantly when $\hat{\gamma}(\lambda, k)$ was smoothed. The GAi algorithm, based on instantaneous measurements of γ and ξ, performed the worst in all conditions. This is not surprising given that the instantaneous values of γ and ξ vary dramatically from frame to frame causing in turn high levels of musical noise (Cappe, 1994) resulting from rapid fluctuations (over time) of the gain function. This outcome suggests that the smoothing of γ and ξ is necessary to obtain high quality speech free of musical tones. The pattern in performance was very similar when the algorithms were evaluated using the LLR objective measure (Table 3).

The performance of the MMSE algorithm was significantly better ($p < 0.05$) than the GA algorithm in most conditions, except in babble at 0 and 5 dB SNR (see Table 2). Smoothing of γ in the MMSE algorithm yielded a decrement in performance in all conditions (see MMSEsm entries in Table 2). Computationally, the GA algorithm has the advantage over the MMSE algorithm in that its implementation only requires a few multiply and add operations (see Eq. (19)). The MMSE algorithm, on the other hand, requires implementations of Bessel functions or alternatively requires sufficient storage for two-dimensional (γ and ξ) look-up tables. Computationally simple implementations of the MMSE algorithm were reported in Wolfe and Godsill (2001).

Fig. 8 shows spectrograms of an example sentence processed by the subtractive (SS and SSsm) and GA algorithms ($\beta = 0.6$) in 5 dB SNR (babble). It is clear that the GA algorithm yielded significantly lower residual noise than the spectral subtractive algorithms. Informal listening tests confirmed that the GA-enhanced speech signal had a smoother background with no audible musical noise, at least for the SNR levels and types of noise tested. As mentioned earlier (see Section 3), we believe that the GA algorithm does not have musical noise because it inherits some

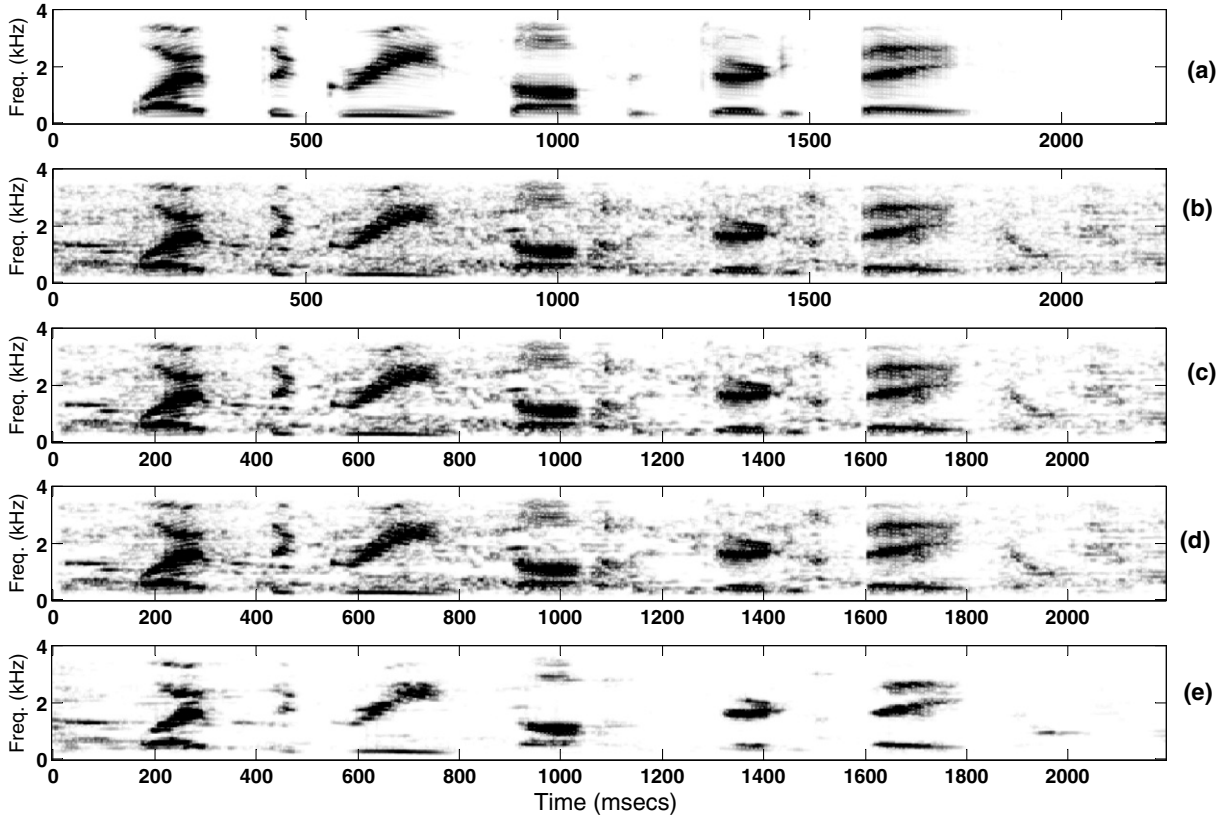

Fig. 8. Spectrograms of the IEEE sentence "Wipe the grease off his dirty face" (sp05.wav in NOIZEUS database) in 5 dB SNR babble processed by the proposed GA algorithm (bottom Panel e), the spectral subtractive algorithm (Panel c) and smoothed spectral subtractive algorithm (Panel d). Top two panels show the spectrogram of the sentence in quiet and in noise, respectively.

of the properties of the MMSE algorithm. In contrast, the spectral subtractive algorithms yielded large amounts of musical noise[3].

6. Conclusions

The present paper presented a new approach (GA algorithm) to spectral subtraction based on geometric principles. Unlike the conventional power spectral subtraction algorithm which assumes that the cross terms involving the phase difference between the signal and noise are zero, the proposed algorithm makes no such assumptions. This was supported by error analysis that indicated that while it is safe to ignore the cross terms when the spectral SNR is either extremely high or extremely low, it is not safe to do so when the spectral SNR falls near 0 dB. A method for incorporating the cross terms involving phase differences between the noisy (and clean) signals and noise was proposed. Analysis of the suppression curves of the GA algorithm indicated that it possesses similar properties as the traditional MMSE algorithm (Ephraim and Malah, 1984). Objective evaluation of the GA algorithm showed that it performed significantly better than the traditional spectral subtraction algorithm in all conditions. Informal listening tests and visual inspection of spectrograms revealed that the GA algorithm had no audible musical noise (at least for the SNR levels tested) and had a smooth and pleasant residual noise. The main conclusion that can be drawn from the present study is that in the context of spectral subtraction algorithms, phase estimation is critically important for accurate signal magnitude estimation. In fact, it is not possible to recover the magnitude spectrum of the clean signal *exactly* even if we had access to the noise signal. Access to phase information is needed.

Acknowledgement

This research was partly supported by Grant No. R01 DC007527 from NIDCD/NIH.

Appendix A

In this appendix, we derive the expression for the cross term error $\varepsilon(k)$ given in Eq. (8). After dividing both sides of Eq. (2) by $|Y(\omega_k)|^2$ we get

$$\varepsilon(k) = \left| 1 - \frac{|X(\omega_k)|^2 + |D(\omega_k)|^2}{|Y(\omega_k)|^2} \right|. \tag{25}$$

After substituting $|Y(\omega_k)|^2$ from Eq. (2) in the above equation, and dividing both numerator and denominator by $(|X(\omega_k)|^2 + |D(\omega_k)|^2)$ we get

$$\varepsilon(k) = \left| 1 - \frac{1}{1 + \frac{\sqrt{\xi(k)}}{\xi(k)+1} 2\cos(\theta_X(k) - \theta_D(k))} \right|, \tag{26}$$

where $\xi(k) \triangleq |X(\omega_k)|^2/|D(\omega_k)|^2$. Finally, after computing the common denominator and simplifying the above equation, we get Eq. (8).

Appendix B

In this appendix, we derive the expressions given in Eqs. (15) and (16) for c_{YD} and c_{XD}. It is easy to show that the following relationships hold:

$$a_X^2 = a_Y^2 + a_D^2 - 2a_Y a_D \cos(\theta_Y - \theta_D), \tag{27}$$

$$a_Y^2 = a_X^2 + a_D^2 + 2a_X a_D \cos(\theta_X - \theta_D). \tag{28}$$

Eq. (27) was derived by applying the Law of Cosines to the triangle shown in Fig. 4. Eq. (28) was derived in the main text and is copied for completeness from Eq. (2). After dividing both sides of the above equations by a_D^2 and using the definitions of γ and ξ given in Eqs. (17), and (18), we get

$$\xi = \gamma + 1 - 2\sqrt{\gamma} \cdot c_{YD}, \tag{29}$$

$$\gamma = \xi + 1 + 2\sqrt{\xi} \cdot c_{XD}. \tag{30}$$

After solving for c_{YD} and c_{XD} in the above equations, we get Eqs. (15) and (16).

It is worth noting here that after using the fact that c_{YD} and c_{XD} are bounded (e.g., $|c_{YD}| \leqslant 1$), we can use Eq. (29) to derive the following bounds on ξ:

$$(\sqrt{\gamma} - 1)^2 \leqslant \xi \leqslant (\sqrt{\gamma} + 1)^2. \tag{31}$$

Restricting ξ to lie within the above range, ensures that $|c_{YD}| \leqslant 1$ and $|c_{XD}| \leqslant 1$.

[3] Audio demonstrations of example sentences enhanced by the GA algorithm can be found at: http://www.utdallas.edu/~loizou/speech/demos/. MATLAB code is available from the second author.

References

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

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (2), 113–120.

Cappe, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. 2 (2), 346–349.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6), 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 Acoustics, Speech, Signal Processing, Vol. I. pp. 145–148.

Hirsch, H., Pearce, D., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ISCA ITRW ASR200.

Hu, Y., Loizou, P., 2007. Subjective evaluation and comparison of speech enhancement algorithms. Speech Communication 49, 588–601.

Hu, Y., Loizou, P., 2006. Evaluation of objective measures for speech enhancement. In: Proc. Interspeech, pp. 1447–1450.

Hu, Y., Loizou, P., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Language Processing 16 (1), 229–238.

ITU, 2000. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation 862.

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

Kitaoka, N., Nakagawa, S., 2002. Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task. In: Proc. Internat. Conf. Spoken Language Processing, pp. 477–480.

Lockwood, P., Boudy, J., 1992. Experiments with a Non-linear Spectral Subtractor (NSS) Hidden Markov Models and the projections for robust recognition in cars. Speech Commun 11 (2–3), 215–228.

Loizou, P., 2005. Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum. IEEE Trans. Speech Audio Process. 13 (5), 857–869.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press LLC, Boca Raton, FL.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512.

Papoulis, A., Pillai, S., 2002. Probability Random Variables and Stochastic Processes, fourth ed. McGraw-Hill, Inc., New York.

Vary, P., 1985. Noise suppression by spectral magnitude estimation: Mechanism and theoretical limits. Signal Process. 8, 387–400.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process. 7 (3), 126–137.

Weiss, M., Aschkenasy, E., Parsons, T., 1974. Study and the development of the INTEL technique for improving speech intelligibility. Technical Report NSC-FR/4023, Nicolet Scientific Corporation.

Wolfe, P., Godsill, S., 2001. Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. In: Proc. 11th IEEE Workshop on Statistical Signal Processing, pp. 496–499.

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 (6), 579–582.