# Phase-Based Dual-Microphone Robust Speech Enhancement

Parham Aarabi, *Member, IEEE,* and Guangji Shi

***Abstract*—A dual-microphone speech-signal enhancement algorithm, utilizing phase-error based filters that depend *only* on the phase of the signals, is proposed. This algorithm involves obtaining time-varying, or alternatively, time-frequency (TF), phase-error filters based on prior knowledge regarding the time difference of arrival (TDOA) of the speech source of interest and the phases of the signals recorded by the microphones. It is shown that by masking the TF representation of the speech signals, the noise components are distorted beyond recognition while the speech source of interest maintains its perceptual quality. This is supported by digit recognition experiments which show a substantial recognition accuracy rate improvement over prior multimicrophone speech enhancement algorithms. For example, for a case with two speakers with a 0.1 s reverberation time, the phase-error based technique results in a 28.9% recognition rate gain over the single channel noisy signal, a gain of 22.0% over superdirective beamforming, and a gain of 8.5% over postfiltering.**

***Index Terms*—Microphone arrays, speech processing, speech recognition, time-frequency analysis.**

## I. Introduction

In VARIOUS applications such as, speech recognition and automatic teleconferencing, the recorded speech signals may be corrupted by noises which can include Gaussian noise, speech noise (unrelated conversations), and reverberation [19]. This corruption often degrades the performance of these systems. For example, in speech recognition systems noise results in a lower speech recognition accuracy rate [19], [21]. As a result, various speech enhancement techniques have been investigated in the past [7]–[9], [15], [18], [22].

The fusion of multiple acoustic signals obtained from an array of microphones is a problem that has received much attention in recent years [2], [3], [7], [14], [15]. This can be partly attributed to the fact that such approaches hold the potential for significant noise removal, thereby enabling many applications including robust speech recognition [2], [7]. Examples of multimicrophone techniques include independent component analysis (ICA) [7], [15] and various beamforming algorithms [9], [10], [17], [18], [23].

While many algorithms exist for the uncorrelated noise situation (such as postfiltering with Wiener filters), practical situations often involve correlated noise [17], [21]. In [10], a solution for correlated noises was proposed. The idea was to improve the signal cross spectrum by estimating the noise cross spectrum during silence intervals and subtracting it from the cross power spectrum of the recorded segment.

Another successful speech enhancement technique was discussed in [9]. The superdirective beamformer was shown to have an approximately 20% higher speech recognition accuracy rate than that the previously discussed techniques.

In this paper, a time-frequency filtering technique is proposed that rewards, or punishes, individual TF blocks (i.e., a certain frequency for a given time segment) based on the observed phases and the expected phases of those blocks. This has the aim of maintaining the spectral structure of the speech source of interest, and thereby, the main contents of that speech source, while damaging the spectral contents of other sources, hopefully beyond recognition. This technique, as will be shown, requires knowledge regarding the time difference of arrival (TDOA) of the speech source of interest. Furthermore, it is an ad-hoc technique while its initial formation was obtained as a result of TDOA estimation.

## II. Problem Statement and Prior Work

In this paper, we target the problem of enhancing noisy speech signals recorded by two microphones with known TDOAs. In general, the following dual-microphone can be used:

$$x_1(t) = s(t) * h_{s1}(t) + n_1(t) \tag{1}$$
$$x_2(t) = s(t) * h_{s2}(t) + n_2(t) \tag{2}$$

which can be represented in the frequency domain as

$$X_1(\omega) = S(\omega)H_{s1}(\omega) + N_1(\omega) \tag{3}$$
$$X_2(\omega) = S(\omega)H_{s2}(\omega) + N_2(\omega) \tag{4}$$

where $h_{s1}(t)$ and $h_{s2}(t)$ are the impulse responses associated with the speech source for the first and second microphone, respectively, $x_1(t)$ and $x_2(t)$ are the signals obtained by the microphones, and $s(t)$, $n_1(t)$, and $n_2(t)$ are the main source, and the noise signal associated with each microphone. The goal of speech enhancement is to combine or process the observed signals $x_1(t)$ and $x_2(t)$ in order to obtain a perceptually equivalent version of $s(t)$. In this effort, a variety of techniques have been proposed, the most common of which is beamforming [9], [18].

### A. Beamforming and Super-Directive Beamforming

We can extend the dual microphone model of (1) and (2) to the $M$ microphone case, as shown

$$\mathbf{x}(t) = \mathbf{h}(t) * s(t) + \mathbf{n}(t) \tag{5}$$

Manuscript received October 3, 2003; revised January 13, 2004. This paper was recommended by Associate Editor S. Singh.

The authors are with the Edward S. Rogers, Sr., Department of Electrical and Computer Engineering, University of Toronto, Toronto, M5S364 ON, Canada (e-mail: parham@ecf.utoronto.ca).

Digital Object Identifier 10.1109/TSMCB.2004.830345

where $s(t)$ is the speech signal of interest at time t, the microphone signal vector $\mathbf{x}(t)$ is a column vector containing the $M$ microphone signals at time $t$; $\mathbf{h}(t)$ is a column vector of the impulse responses of each microphone for the given source of interest; and $\mathbf{n}(t)$ is a vector of possibly dependent noises. In practice, we must sample a finite segment of the microphone signals. Assuming that we take an $N$ sample segment (with sampling rate $F_s$) and take its Fourier transform, (5) can be restated in the frequency domain as

$$\mathbf{X}(\omega) = \mathbf{H}(\omega)S(\omega) + \mathbf{N}(\omega) \tag{6}$$

where the capital letters are all discrete-time Fourier transforms of their lower-cased time domain representations. Note that because we are taking Fourier transforms of finite signal segments (which according to the DFT or periodically extended), our representation in the frequency domain is nonzero at discrete values (i.e., $\omega$ is defined at a discrete set of values starting from 0 and incrementing or decrementing in $2\pi F_s/N$ steps). While we have used a general impulse response model in (5), we assume that the TDOAs relative to the first microphone for the speech signal of interest are known. In such a scenario, the beamforming operation can be defined as

$$\tilde{X}(\omega) = \mathbf{A}(\omega)\mathbf{X}(\omega) \tag{7}$$

where $\mathbf{A}(\omega)$ is a row of complex weights defined as follows [9]:

$$\mathbf{A}(\omega) = \frac{\mathbf{\Gamma}^{-1}(\omega)\mathbf{d}(\omega)}{\mathbf{d}^{*}(\omega)\mathbf{\Gamma}^{-1}(\omega)\mathbf{d}(\omega)} \tag{8}$$

and the steering vector $\mathbf{d}(\omega)$ is defined as

$$\mathbf{d}(\omega) = [1, e^{-j\omega\tau_2}, \ldots, e^{-j\omega\tau_M}] \tag{9}$$

where $\tau_2, \tau_3, \ldots, \tau_M$ are the set of TDOAs for the second to $M$th microphones relative to the first microphone and corresponding to the position of the sound source of interest. Finally, the coherence matrix $\mathbf{\Gamma}(\omega)$ is defined as

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & \Gamma_{X_1X_2} & \cdots & \Gamma_{X_1X_M} \\ \Gamma_{X_2X_1} & 1 & \cdots & \Gamma_{X_2X_{M-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{X_MX_1} & \Gamma_{X_MX_3} & \cdots & 1 \end{pmatrix}. \tag{10}$$

For delay-and-sum beamforming, we have: $\Gamma_{X_uX_v}(\omega) = 0$, for $u \neq v$.

For superdirective beamforming, we have [9]

$$\Gamma_{X_uX_v}(\omega) = \frac{\sin\left(\frac{\omega d_{uv}}{c}\right)}{\frac{\omega d_{uv}}{c}\left(1 + \frac{\sigma_n^2}{P_{NN}(\omega)}\right)} \tag{11}$$

where $c$ is the speed of sound, $d_{uv}$ is distance between the $u$th and $v$th microphones, $\sigma_n^2$ is the variance of uncorrelated sensor noise, and $P_{NN}(\omega)$ is the power spectral density of the diffuse noise field. As suggested in [9], a $-20$ dB to $-40$ dB sensor noise to room noise ratio gives good results in most practical situations.

### B. Postfiltering

Another widely used array speech enhancement technique is postfiltering [10], [17]. A postfilter consists of a beamformer followed by a time varying filter.

Based on the Wiener–Hopf equation, the transfer function of the Wiener filter at any frequency $\omega$ is expressed as

$$W(\omega) = \frac{\Phi_{\tilde{X}S}(\omega)}{\Phi_{\tilde{X}\tilde{X}}(\omega)} \tag{12}$$

where $\Phi_{\tilde{X}\tilde{X}}(\omega)$ is the power spectral density (PSD) of the beamformer output and $\Phi_{\tilde{X}S}(\omega)$ is the cross power spectral density (CSD) of output and the original clean signal of interest. The beamformer output $\tilde{X}(\omega)$ is defined as $\tilde{X}(\omega) = X_1(\omega) + X_2(\omega)e^{j\omega\tau}$, where $\tau$ is the time difference of arrival between the two microphones (assumed to be known), and $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the signals recorded by the two microphones. In [17], the performance of the postfilter was analyzed in detail. It was shown that the following realization of the Wiener filter, shown here for the case of two microphones, gives good results

$$W_1(\omega) = \frac{\Phi_{X_1(X_2e^{j\omega\tau})}(\omega)}{\frac{1}{2}\left(\Phi_{X_1X_1}(\omega) + \Phi_{X_2X_2}(\omega)\right)} \tag{13}$$

where $\Phi_{X_1(X_2e^{j\omega\tau})}(\omega)$ is the CSD of the time-aligned input signals from the different microphones, and, $\Phi_{X_1X_1}(\omega)$ and $\Phi_{X_2X_2}(\omega)$ are the PSDs for signals of the first and second microphones, respectively. Equation (13) is a good approximation of (12) under the assumptions that noise at each sensor is uncorrelated and there is no correlation between noise and the desired signal. When there are more microphones, [17] showed it is better to use a directivity-controlled array rather than a conventional beamformer [17]. In this paper, we will concentrate on the two-microphone case only, and hence, will only consider the postfilter of (13).

## III. Preliminaries

The time-domain representation of speech signals often fails to convey clear information regarding the contents of the speech signal. The frequency domain representation of speech, however, illustrates the harmonics and formants which are essential to the recognition of speech [19]. In order to visualize the formants of a speech signal, a short-time Fourier transform representation is required since the formants change with respect to time.

Assuming that we have a recorded speech signal $x(t)$, we sample it with sampling frequency $F_s$ resulting in the discrete signal $\hat{x}(n) = x(nT_s)$, where $T_s = 1/F_s$ is the sampling period. We partition $\hat{x}(n)$ into half overlapping $N$-sample segments which are windowed by a Hanning (or more correctly, Von Hann) window (the windowing function should be chosen such that the original time-domain signal can be obtained by overlapping and adding the windowed segments). We define the Fourier transform of the $k$th time segment as $X_k(\omega)$, where, as before, the frequency index $\omega$ is defined at discrete frequency values (in steps of $2\pi F_s/N$) due to the finite time-window.

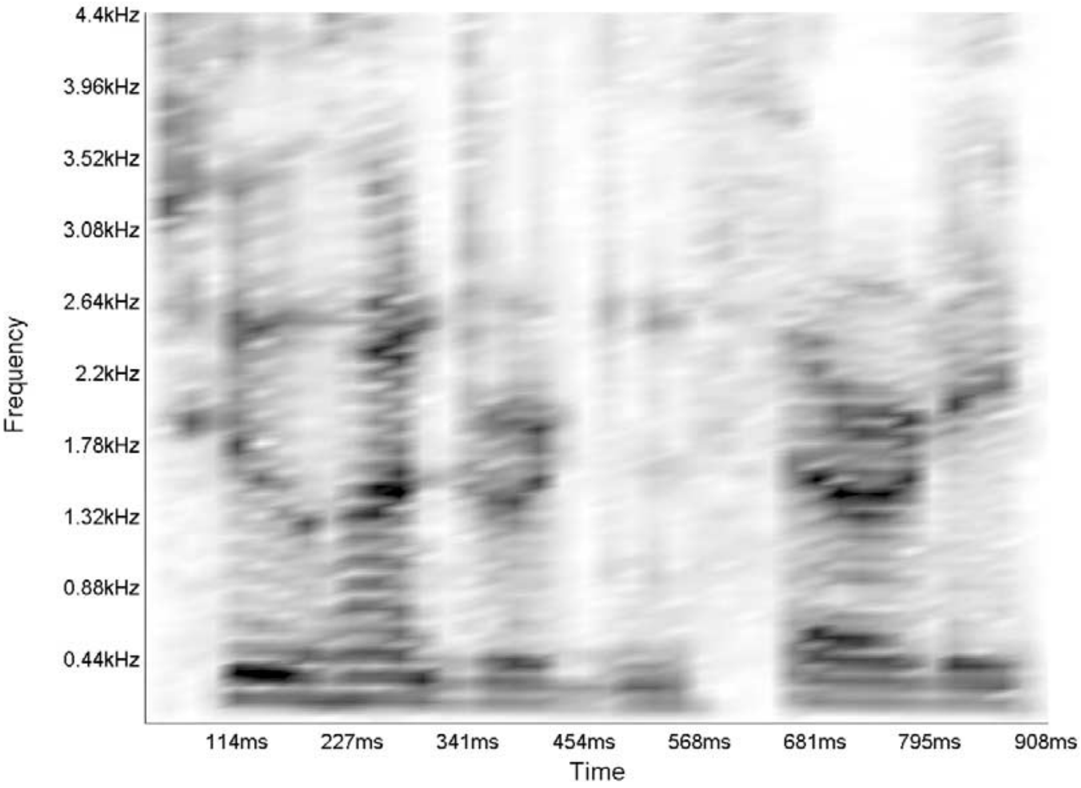

Fig. 1. Spectrogram of the phrase "two-hundred and thirty."

$X_k(\omega)$ can be viewed as the discrete time-frequency (TF) transformation of $x(t)$. Note that while $X_k(\omega)$ is directly obtained from $x(t)$, the inverse [i.e., obtaining $x(t)$ back from $X_k(\omega)$] can be done by taking the inverse fast Fourier transform (IFFT) of $X_k(\omega)$ for each segment, overlapping and adding the segments, and reconstructing the continuous signal from the discrete-time signal. It should be mentioned that although there are mathematically $N$ discrete frequency points, only half (i.e., $N/2$) are of value since the remaining half are just the complex conjugates of the first half (due to the fact that the time-domain signal is real). As a result, we will just analyze $N/2$ discrete frequencies.

As a result, after the TF transformation, we have a complex TF image, or a set of phase and magnitude TF images. The magnitude image is known as the spectrogram, and is often used to depict the characteristics of difference vowels. Fig. 1 shows the spectrogram for the male utterance of the phrase "two-hundred and thirty." While spectrograms are often relied upon to convey information about the contents of the speech signal, their phase counterparts are not often used. Although the phase information of speech signals has not been fully exploited in most speech recognition systems, it is almost always used in multi-microphone settings for time-delay estimation.

In this paper, we often use the term "time-frequency block" to correspond to a certain frequency component for a certain time segment. This is in fact equivalent to a short-time DFT over multiple segments. However, we will use the time-frequency notation since our short time filtering strategy (alternatively known as time-frequency masking in the literature) is a time-varying filter.

## IV. Phase-Based Time-Frequency Masking

In this section, we will introduce the concept of phase dependent time-frequency masking for the purpose of speech signal enhancement. The basic premise is that working in the time-frequency domain can result in better speech signal enhancement than either the time or the frequency domain techniques.

### *A. TDOA Estimation and Its Relation to Phase Error Minimization*

Note that TDOA estimation is not the focus of this paper. In fact, the proposed algorithm and the corresponding experiments utilize prior knowledge regarding the TDOA of the speaker of interest (i.e., no online TDOA estimation is made). Nevertheless, TDOA estimation does provide a unique insight and interesting introduction to the proposed time-frequency reward-punish algorithm. This, and only this, is the reason behind the inclusion of this section.

Assuming two microphones are present in an environment with a sound source, the sound waves which are produced by the source will arrive at the two sources at different times. Since sound travels at a speed of approximately 345 m/s in air, the time differences, known as the time-difference of arrivals, will be small (about 1–2 ms) compared to the length of the speech segments used in the spectrograms (typically 20 ms long). As a result, the spectrogram for the two channels will be very similar. The time-frequency phase images, however, will be quite different because of the time delay between the two channels and can be used to estimate the TDOA of the sound signal between the two microphones.

There are many different algorithms that attempt to estimate the most likely TDOA between a pair of microphones [5], [11], [12]. Usually, these algorithms have a heuristic measure that estimates the likelihood of every possible TDOA, and selects the most likely value. The most widely used TDOA estimator is the generalized cross correlation class, which attempts to filter the cross correlation between two received signals in an optimal or suboptimal manner, and then selects the time index of the peak of the result to be the TDOA estimate. Considering a simplified model of (1) and (2), we have

$$x_1(t) = h_1 \cdot s(t) + n_1(t) \tag{14}$$

$$x_2(t) = h_2 \cdot s(t-\tau) + n_2(t). \tag{15}$$

The two microphones receive a time-delayed and scaled version of the source signal $s(t)$ without modeling reverberations. The goal of TDOA estimation is to estimate $\tau$ given the microphone signals $x_1(t)$ and $x_2(t)$, where we assume $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of these signals, respectively. The most common solution to this problem is the generalized cross correlation approach shown below [12]

$$\tilde{\tau} = \arg\max_{\beta} \int_{-\infty}^{\infty} W(\omega) X_1(\omega) \overline{X_2(\omega)} e^{-i\omega\beta} dw \tag{16}$$

where $\tilde{\tau}$ is an estimate of the original source signal delay between the two microphones. The above equation assumes that there is only one segment of the microphone signals available. In reality, in order to make sure that the speech source is stationary within a single segment, only 10–20 ms segments can be used. Since there may be many such segments available, a different form of (16) that incorporates N different signal segments is shown

$$\tilde{\tau} = \arg\max_{\beta} \sum_{k=1}^{N} \int_{-\infty}^{\infty} W(k,\omega) X_{1,k}(\omega) \overline{X_{2,k}(\omega)} e^{-i\omega\beta} dw \tag{17}$$

which, in practice, can be written in the following discrete-frequency form:

$$\tilde{\tau} = \arg\max_{\beta} \sum_{k=1}^{N} \sum_{\omega=-\omega_s}^{\omega_s} W(k,\omega) X_{1,k}(\omega) \overline{X_{2,k}(\omega)} e^{-i\omega\beta} \tag{18}$$

where $\omega_s$ is the highest frequency of interest in radians. The above two equations assume that the sound source time delay between the microphones remains constant for all of the signal segments. The actual choice of the weighing function $W(k,\omega)$ has been studied at length for general sound and speech sources. Three different choices, the maximum likelihood (ML) [12], [16], the phase transform (PHAT) [12], [20], and the unfiltered cross correlation (UCC) [13] are

$$W_{ML}(k,\omega) = \frac{|X_{1,k}(\omega)|\,|X_{2,k}(\omega)|}{|N_{1,k}(\omega) \cdot X_{2,k}(\omega)|^2 + |N_{2,k}(\omega) \cdot X_{1,k}(\omega)|^2} \tag{19}$$

$$W_{PHAT}(k,\omega) = \frac{1}{\left|X_{1,k}(\omega) \cdot \overline{X_{2,k}(\omega)}\right|} \tag{20}$$

$$W_{UCC}(k,\omega) = 1. \tag{21}$$

The maximum likelihood weights require knowledge about the spectrum of the microphone dependent noises. The phase transform does not require this knowledge, and hence, has been employed more often due to its simplicity. The unfiltered cross correlation does not utilize any weighing function.

With the PHAT weights, the discrete-frequency cross-correlation reduces to

$$\tilde{\tau}_{PHAT} = \arg\max_{\beta} \sum_{k=1}^{N} \sum_{\omega=-\omega_s}^{\omega_s} \cos\left(\theta_{\beta,k}(\omega)\right) \tag{22}$$

where $\theta_{\beta,k}(\omega)$ is defined as

$$\theta_{\beta,k}(\omega) = \angle X_{1,k}(\omega) - \angle X_{2,k}(\omega) - \omega\beta. \tag{23}$$

Equation (22) involves a maximization that will be achieved when the appropriate choice of $\beta$ equals the TDOA $\tau$, resulting in a decreased phase error for most frequencies. As a result, the phase transform can be (approximately) represented as a phase error minimization technique, which can be defined as [12]

$$\tilde{\tau}_{PHAT} = \arg\min_{\beta} \psi_{\beta} \tag{24}$$

where $\psi_\beta$ is the following phase variance corresponding to the TDOA $\beta$:

$$\psi_\beta = \sum_{k=1}^{N} \sum_{\omega=-\omega_s}^{\omega_s} \theta_{\beta,k}^2(\omega). \tag{25}$$

At the correct time delay $\tau$, (24) will have a minimized phase variance $\psi_\tau$ of

$$\psi_\tau = \min_{\beta} \sum_{k=1}^{N} \sum_{\omega=-\omega_s}^{\omega_s} \theta_{\beta,k}^2(\omega). \tag{26}$$

Ideally, $\psi_\tau$ should be equal to 0. However, due to the presence of noises, reverberations, and the effects of a finite-duration window, the minimum phase variance, or MPV, will be nonzero.

As an example, consider the case when

$$x_1(t) = s(t) + n_1(t) \tag{27}$$

$$x_2(t) = s(t-\tau) + n_2(t) \tag{28}$$

where $n_1(t)$ and $n_2(t)$ are independent Gaussian signals. Using 2 min of a speech signal obtained from a male speaker with a simulated $\tau$ of 0.11 ms, a simulation was performed in order to illustrate the relationship between the amount of noise present and $\psi_\tau$. As shown in Fig. 2, the MPV decreases as the SNR is increased.

Just like noise, reverberations will also increase the MPV. This is shown in a simulated 7 m by 6 m by 2.5 m environment, where the microphone pair and the speech source are positioned as shown in Fig. 3.

By simulating reverberation times between 0 s and 0.4 s, and using Eyring's formula to obtain wall reflection ratios between 0 and 0.87 [12], the MPV versus reverberation time graphs shown in Fig. 4. Note that once the wall reflection ratio was obtained from the reverberation time, the Image model technique [6] up

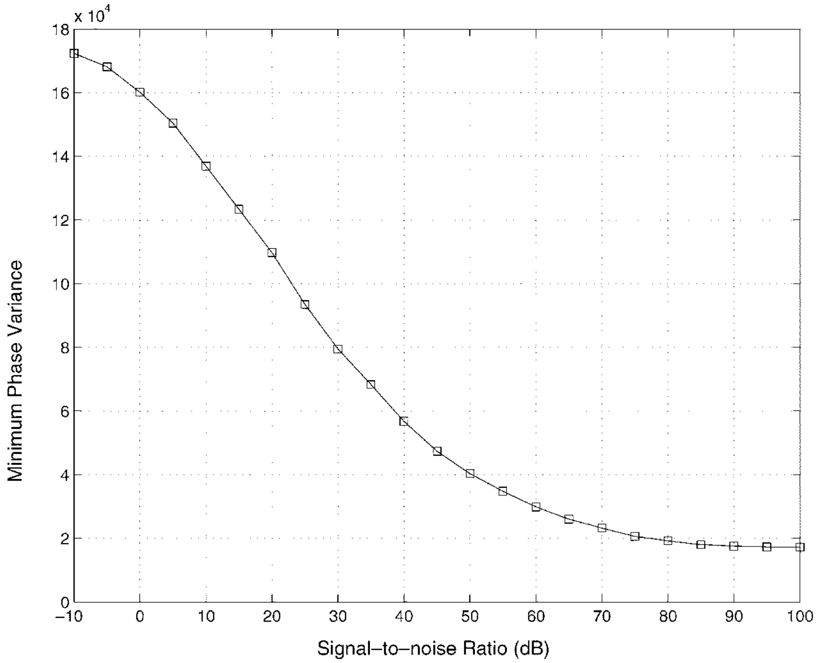

Fig. 2. Relation between the MPV and the SNR.

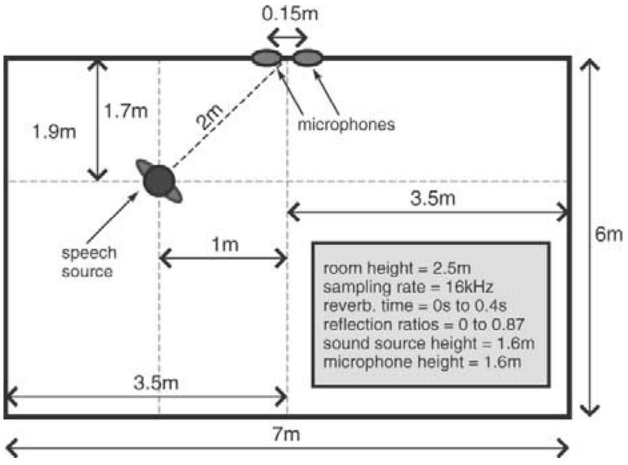

Fig. 3. Simulated reverberant environment.

to ninth-order reflections was used to estimate the impulse response for each microphone.

Fig. 4 shows that the greater the reverberations, the higher the MPV. This is analogous to the results of Fig. 2 where the greater the noise, the higher the MPV.

### B. Phase-Error Based Time-Frequency Masking

Clearly, the MPV, which is the sum-squared phase error for all TF blocks, is a good indication of the level of noise or reverberation that is present in the entire signal. Note that in this paper, a TF block corresponds to a specific frequency component of a specific time segment. Since we may have several time segments, indexed by $k$, and several frequencies, indexed by $\omega$, we call this component the $k$th time and the $\omega$th frequency block.

In this paper, we propose that the block-based phase error, $\theta_{\tau,k}(\omega) = \angle X_{1,k}(\omega) - \angle X_{2,k}(\omega) - \omega\tau$, be used as a metric for the amount of noise in an individual block. As shown in the Appendix , the phase error in fact defines an upper bound for the signal to noise ratio of a given TF block.

While we use the definition $\theta_{\tau,k}(\omega) = \angle X_{1,k}(\omega) - \angle X_{2,k}(\omega) - \omega\tau$, it is assumed that this phase error is always wrapped between $-\pi$ to $\pi$.

As a result of the relation between the signal to noise ratio of a given TF block and its phase error, we propose that the phase error be used as a reward-punish criteria for noise removal. In other words, we will punish TF blocks with large phase errors (i.e., scale down their amplitudes) while keeping low phase error blocks intact. However, in order to implement such a phase-error dependent reward–punish algorithm, its aggressiveness and sensitivity to different phase errors must be analyzed.

This scaling can obviously be done for each of the two channels, separately, with the end result of each channel being combined by delay-and-sum beamforming or more elaborate techniques. Throughout this paper, we will use delay-and-sum beamforming as a means of integrating the two masked microphone signals. In the experiments, we will compare the result of performing time-frequency masking on only a single channel, and masking on both channels followed by delay-and-sum beamforming.

### C. Masking Criterion

Aggressive reward-punish techniques work well in low SNR situations and not very well with high SNRs. The opposite is true for less aggressive reward-punish techniques. The reason behind this can be attributed to the fact that aggressive techniques damage the clean signal, thereby limiting the highest possible output SNR, while less aggressive techniques do little harm to the actual signal while having little effect on the noise. The following discussion will analyze this point further:

Assuming that we treat all TF blocks similarly then our goal is to reduce the amplitude of the $k$th time and $\omega$th frequency

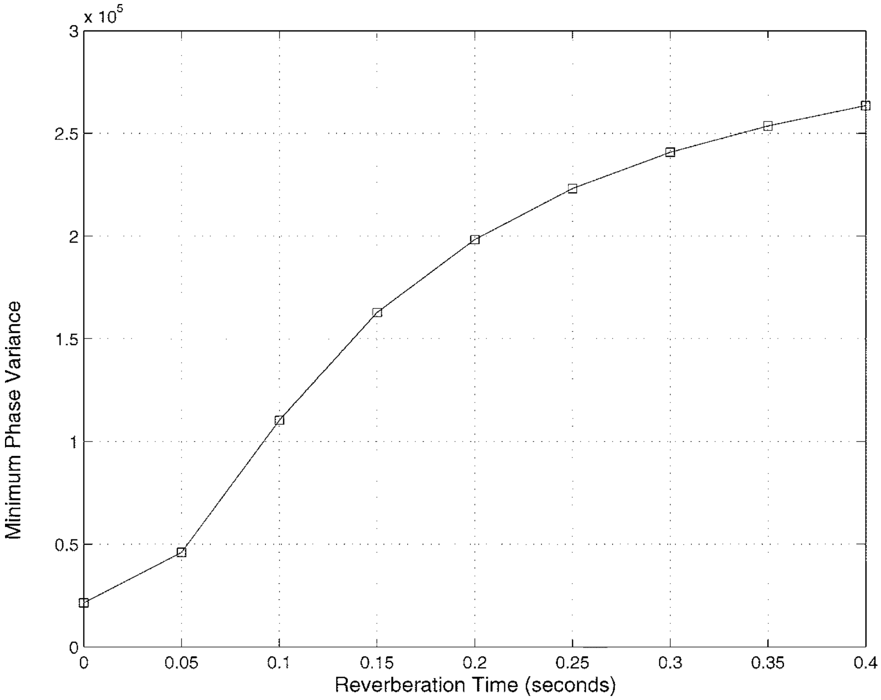

Fig. 4. Relation between the MPV and the reverberation time.

block with a parameter $\eta_k(\omega)$ ranging from 0 to 1. As a result, our overall SNR $\mathbf{R}$ can be defined as

$$\mathbf{R} = \frac{\sum\limits_{\omega}\sum\limits_{k}|S_k(\omega)|^2}{\sum\limits_{\omega}\sum\limits_{k}|N_k(\omega)|^2\eta_k(\omega)^2 + |S_k(\omega)|^2(1-\eta_k(\omega))^2} \quad (29)$$

where $S_k(\omega)$ is the coefficient of the $k$th time block at frequency $\omega$ and $N_k(\omega)$ is the noise component of the same TF block. Note that the above definition of SNR includes damage to the clean signal as additional noise.

Clearly, our goal is to maximize this SNR. Differentiating with respect to $\eta_k(\omega)$ and equating to 0, we obtain the following optimal scaling equation (a.k.a. the Wiener filter):

$$\eta_k(\omega) = \frac{R_k(\omega)}{1+R_k(\omega)} \quad (30)$$

where $R_k(\omega)$ is defined as the SNR of the $\{\omega, k\}$ TF block. In other words

$$R_k(\omega) = \frac{|S_{j,k}(\omega)|^2}{|N_{j,k}(\omega)|^2}. \quad (31)$$

Equation (30) defines an optimal amplitude scaling factor that would maximize the overall SNR given knowledge about the SNR of each individual block. The remaining challenge now is to estimate this block-SNR.

By using the SNR bound of the Appendix and the optimal block scaling defined by (30), we obtain the following block scaling bound:

$$\eta_k^*(\omega) \leq \frac{1}{1+\sin^2\left(\frac{\theta_{\tau,k}(\omega)}{2}\right)} \quad (32)$$

where $\eta_k^*(\omega)$ is the upper bound for the scaling of a given TF block.

In this paper, we propose that the following equation be used as a parameterized scaling strategy for each TF block:

$$\eta_k(\omega) = \frac{1}{1+\gamma\theta_{\tau,k}^2(\omega)} \quad (33)$$

where $\eta_k(\omega)$ is the TF block attenuation function and $\gamma$ is a fixed constant.

A simulation was performed on 100 12-ms male speech segments consisting of both voiced and unvoiced segments (about 80% voiced and 20% unvoiced) corrupted by white Gaussian noise in order to analyze the effectiveness of block scaling strategy. Fig. 5 illustrates the boundary scale as a function of the phase error and compares it with the scaling of (33) with parameters $\gamma = 1$ and $\gamma = 5$.

Fig. 6 illustrates the performance of the TF block scaling at the bound scales defined by (32) for the setup of Fig. 3 without reverberations and with independent Gaussian noise added to each channel. Note that the point of these simulations is not to fully test the algorithm but just show the relative effects of different filter choices. For a detailed set of SNR-based simulations, please refer to [4]. As shown, the SNR improvement is not very large (i.e., is less than 3 dB).

Fig. 7 illustrates the SNR improvement obtained by the proposed technique using the attenuation function of (33) and with several different values for $\gamma$.

As shown in Fig. 7, higher values of $\gamma$ punish high phase-error blocks more severely, resulting in improved performance in low SNR situations and worse performance in high SNR situations. As a result, a value of 5 for $\gamma$ was chosen and used for the experiments in this paper as a result of its consistent SNR gains over a broad range of input SNRs.

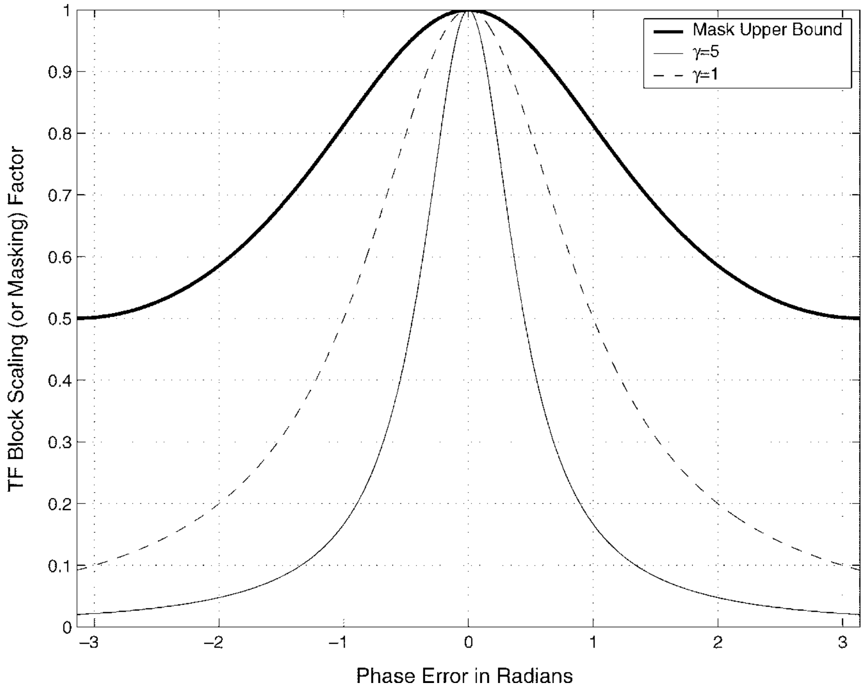

Fig. 5. Comparison of the TF block scale bound defined by (32) and the scaling proposed by (33).

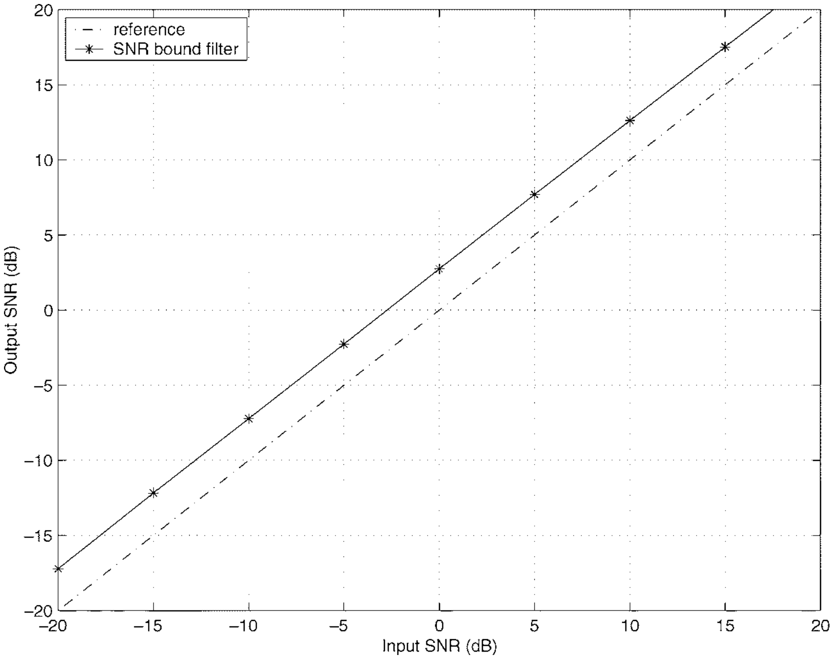

Fig. 6. Performance of the maximum allowable block scaling defined by (32).

### *D. Relation Between Phase-Error-Based Masking and Postfiltering*

The postfilter suggested by (13) is in fact related to the phase-error based filtering proposed in this paper. The postfilter suggested by [17], which was shown in (13), can be approximated (using single-segment CSD and PSD estimates) as

$$W_1(\omega) \approx \frac{\mathrm{Re}\left(X_1(\omega)\overline{X_2(\omega)e^{j\omega\tau}}\right)}{\frac{1}{2}\left(|X_1(\omega)|^2+|X_2(\omega)|^2\right)} = \frac{X_1(\omega)\overline{X_2(\omega)}e^{-j\omega\tau}+\overline{X_1(\omega)}X_2(\omega)e^{j\omega\tau}}{\left(|X_1(\omega)|^2+|X_2(\omega)|^2\right)} \quad (34)$$

which can be simplified to

$$W_1(\omega) = \frac{2|X_1(\omega)||X_2(\omega)|\cos(\theta_\tau(\omega))}{|X_1(\omega)|^2+|X_2(\omega)|^2} = \frac{2\cos(\theta_\tau(\omega))}{\frac{|X_1(\omega)|}{|X_2(\omega)|}+\frac{|X_2(\omega)|}{|X_1(\omega)|}} \quad (35)$$

where $\theta_\tau(\omega)$ is the phase-error for the current time segment, as defined previously.

As shown in Fig. 8, when $\theta_\tau(\omega)$ is small, the filter has a value close to 1, and when $\theta_\tau(\omega)$ is large, the filter has a small value. Hence, the postfilter of [17] (or, at least, a single-segment based CSD and PSD estimate version of it) is in fact a special case of the phase-error based filtering discussed in this paper.

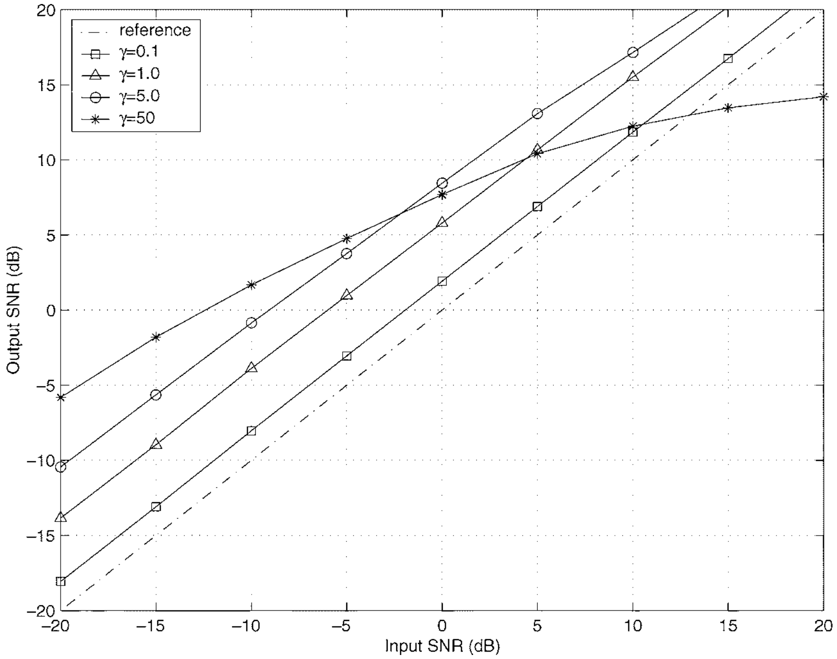

Fig. 7. SNR ratio improvement using the attenuation function of (33).

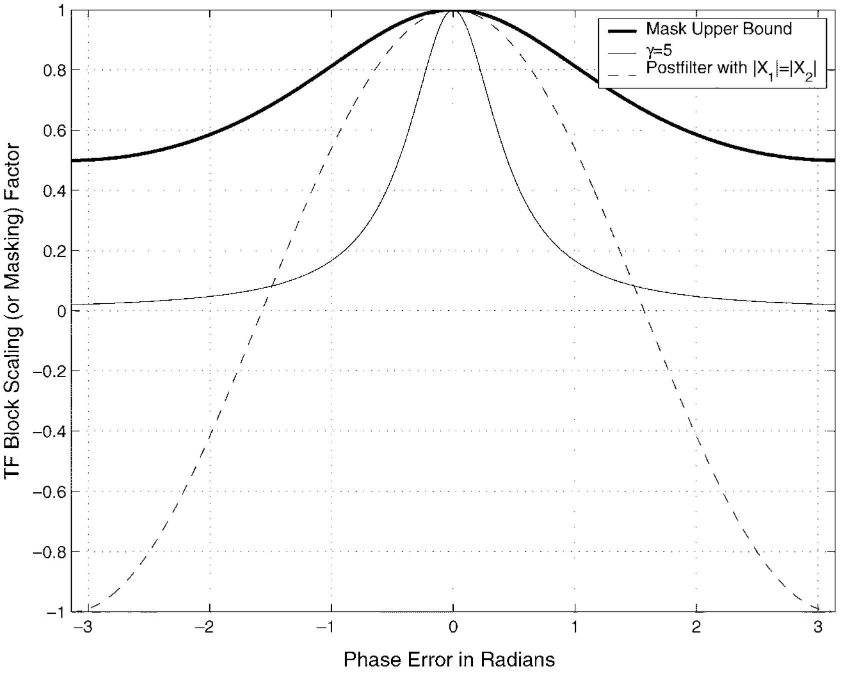

Fig. 8. Illustration of the relation between postfiltering at the proposed phase-error-based masking approach.

Two problems with the postfilter can be observed from Fig. 8. First, the postfilter has a width which can make it too lenient on high phase-error frequencies or TF blocks, unlike the filter proposed in this paper with $\gamma = 5$. Second, for high phase-errors, the postfilter can leave the magnitude of the block intact and just flip the sign (i.e., this corresponds to a multiplication by $-1$). This, in effect, leaves some noisy TF blocks or frequencies in the postfilter output. The proposed filter does not suffer from this problem.

## V. Experimental Evaluation

While SNR gain simulations are useful, they cannot truly convey the effectiveness (or lack thereof) of a speech enhancement technique. A much better test is the performance of the enhanced speech signal (using both the proposed technique and alternative techniques) on a speech recognition system.

In this section, an experiment was conducted with five different speakers. A speaker-independent single-digit recognition system (with no training) was built based on the voice extreme module from Sensory Inc. Detailed information about the architecture of this recognition system can be found in [1]. This module was programmed to only recognize a set of ten preset digits. While the performance of the speaker-independent digit recognition modules may not be on par with more capable speech recognition systems, their portability and universal applicability made them an attractive platform for testing different algorithms. Furthermore, since we care only a perceptual

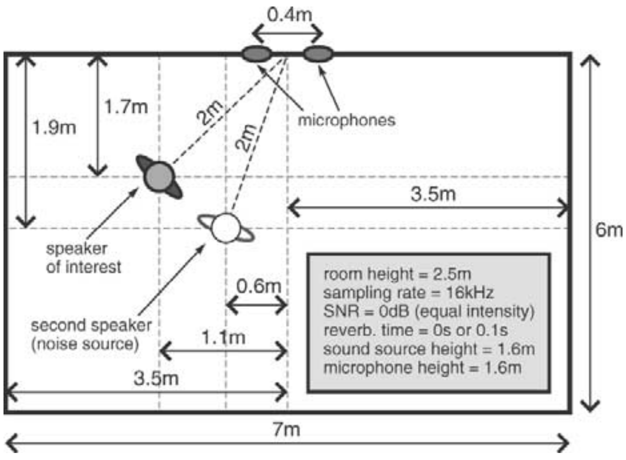

Fig. 9. Simulation setup with a speech source of interest, an unwanted speech source, and two microphones.

TABLE I
DIGIT RECOGNITION ACCURACY RATE COMPARISON (DS=DELAY-AND-SUM BEAMFORMING, SD=SUPERDIRECTIVE BEAMFORMING, PEF1=PHASE-ERROR BASED FILTERING APPLIED TO ONLY THE SIGNAL OF THE FIRST (LEFT) MICROPHONE, PEF2=PHASE-ERROR BASED FILTERING APPLIED TO ONLY THE SIGNAL OF THE SECOND (RIGHT) MICROPHONE, PEF+DS=PHASE-ERROR BASED FILTERING ON BOTH SIGNALS FOLLOWED BY DELAY-AND-SUM BEAMFORMING, PF=POSTFILTERING, CLEAN= ORIGINAL SIGNAL BEFORE THE ADDITION OF SECOND SPEAKER NOISE, NOISY=UNALTERED NOISY SIGNAL OF THE FIRST (LEFT) MICROPHONE) WITH TWO MICROPHONES AT 0 dB

| | S1 | S2 | S3 | S4 | S5 | Average |
|---|---|---|---|---|---|---|
| Clean | 78.4% | 94.3% | 88.6% | 86.3% | 86.3% | 86.8% |
| Noisy | 25.0% | 33.0% | 37.5% | 67.5% | 68.8% | 46.4% |
| DS | 33.0% | 56.8% | 58.0% | 77.5% | 67.5% | 58.6% |
| SD | 36.4% | 61.4% | 60.2% | 75.0% | 68.8% | 60.4% |
| PF | 59.1% | 81.8% | 80.7% | 87.5% | 82.5% | 78.3% |
| PEF1 | 80.7% | 89.8% | 89.8% | 92.5% | 75.0% | 85.6% |
| PEF2 | 80.7% | 85.2% | 90.9% | 93.8% | 76.3% | 85.4% |
| PEF + DS | 83.0% | 80.7% | 88.6% | 91.3% | 78.8% | 84.5% |

analysis of the different algorithms, the Sensory, Inc. module was deemed to be acceptable for such a task.

The speech recognition system, which is small enough to be embedded in handheld applications, recorded 20–30 random digits (in the 0–9 range with a corresponding ten digit vocabulary) from each speaker. The noise was artificially-added speech noise consisting of a male conversation, resulting in a signal to noise ratio of 0 dB. The sampling frequency was 16 KHz. For the phase-error TF mask, (17) was used with $\gamma = 5$. The setup of the experiment is shown in Fig. 9. Two cases were considered, one without reverberation, and one with a reverberation time of 0.1 s.

### A. Experiment Without Reverberation

Table I shows the digit recognition rate test for five different speakers. For consistency, each simulated experiment was conducted four separate times with different secondary speech signals (i.e., the noise signal), and the results of the four trials were averaged. The output from the phase-error filter applied to either channel (PEF1 or PEF2) gives the highest average recognition accuracy, which is very close to the average recognition accuracy rate of the clean signal. Both the delay-and-sum beamformer and the superdirective beamformer are able to obtain small recognition accuracy rate gains (12.2% for delay-and-sum beamforming, 14% for superdirective beamforming). Postfiltering (using the approach of [17]) does better than conventional beamforming, with a gain that is 6.2% less than that of phase-error filtering with delay-and-sum beamforming, and is an improvement of 31.9% over the noisy signal. Finally, the PEF+DS technique proposed in this paper (with $\gamma = 5$) results in an improvement of 38.1% over the single microphone noisy signal. This result is much better than the other techniques under analysis. Note that all of these results have been obtained without taking reverberation into account and only used two microphones.

TABLE II
DIGIT RECOGNITION ACCURACY RATE COMPARISON (DS=DELAY-AND-SUM BEAMFORMING, SD=SUPERDIRECTIVE BEAMFORMING, PEF1=PHASE-ERROR BASED FILTERING APPLIED TO ONLY THE SIGNAL OF THE FIRST (LEFT) MICROPHONE, PEF2=PHASE-ERROR BASED FILTERING APPLIED TO ONLY THE SIGNAL OF THE SECOND (RIGHT) MICROPHONE, PEF+DS=PHASE-ERROR BASED FILTERING ON BOTH SIGNALS FOLLOWED BY DELAY-AND-SUM BEAMFORMING, PF=POSTFILTERING, CLEAN=ORIGINAL SIGNAL BEFORE THE ADDITION OF SECOND SPEAKER NOISE, NOISY=UNALTERED NOISY SIGNAL OF THE FIRST (LEFT) MICROPHONE) WITH 2 MICROPHONES AT 0 dB, INCLUDING A 0.1 s REVERBERATION TIME

| | S1 | S2 | S3 | S4 | S5 | Average |
|---|---|---|---|---|---|---|
| Noisy | 19.3% | 31.8% | 37.5% | 61.3% | 65.0% | 43.0% |
| DS | 31.8% | 38.6% | 51.1% | 57.5% | 70.0% | 49.8% |
| SD | 33.0% | 37.5% | 46.6% | 61.3% | 71.3% | 49.9% |
| PF | 47.7% | 65.9% | 68.2% | 68.8% | 66.3% | 63.4% |
| PEF1 | 69.3% | 71.6% | 68.2% | 76.3% | 80.0% | 73.1% |
| PEF2 | 60.2% | 68.2% | 70.5% | 77.5% | 67.5% | 68.8.% |
| PEF + DS | 63.6% | 73.9% | 69.3% | 81.3% | 71.3% | 71.9% |

It is interesting to note that if the phase-error filter was *only* applied to the signal of the first microphone (PEF1) or the second microphone (PEF2), then the result would be slightly better than the case were both channels are filtered followed by delay-and-sum beamforming (PEF+DS). This suggests that the recognition rate gains are mainly as a result of phase-error filtering and not as a result of delay-and-sum beamforming.

### B. Experiment With Reverberation

We now consider the case were reverberations are present and modeled using the image model technique [6]. The wall reflection ratio was estimated to be 0.57 using Eyring's formula [12], the assumed room dimensions, and the assumed reverberation time of 0.1 s. The digit recognition results, which are shown in Table II, are similar to the case without reverberations. The PEF1 (phase-error based filter applied to only the first channel) technique is a 9.7% higher recognition rate than postfiltering, and 30.1% higher than the noisy signal. The PEF2 (phase-error based filter applied to only the second channel), on the other hand, has only a 5.4% recognition rate gain over postfiltering, and a gain of 25.8% over the noisy signal's recognition rate. Again, for consistency, each simulated experiment was conducted four separate times with different secondary speech signals (i.e., the noise signal), and the results of the four trials were averaged. In all four trials, the variation of the average recognition rate was within ±1%.

It is interesting to note that the combined PEF+DS technique, in the reverberant case, actually results in a slightly lower recognition rate than PEF applied to only the first channel. This corresponds to a gain of 8.5% over postfiltering, and an overall gain of 28.9% over the noisy signal. In this case, it is likely that the lower perceptual quality of the second filtered signal (PEF2) resulted in the lower recognition rate of the combined signals (PEF+DS).

## VI. Conclusions

A phase-error-based time-frequency masking technique was proposed. It was shown through speaker independent digit recognition experiments that the proposed technique achieves a higher digit recognition rate than prior speech enhancement techniques, which include superdirective beamforming, delay-and-sum beamforming, and postfiltering. This higher recognition rate result is consistent in both reverberant and nonreverberant situations.

Further improvements of this work would ideally center in two areas. First, the proposed phase-error filtering technique is ad-hoc, and could greatly improve from a more careful investigation of the phase-error filter shape. Second, it is not clear how this technique would be extended do multiple microphones (as compared to beamforming or postfiltering where the extension is clear). As a result, analyzing the method of application of the algorithm to multiple microphones as well as investigating the benefits of such an extension would be fruitful directions of future research.

## Appendix
## Relation Between Phase Error and SNR

Intuitively, it would seem that a larger phase error for a given TF block is indicative of a lower SNR, and vice versa. We can show this relationship directly as follows.

Assuming that the noise for each given TF block has a constant amplitude in both channels, we can define the contents of TF block at frequency $\omega$ and time-segment $k$ as

$$X_{1,k}(\omega) = |S_k(\omega)|e^{i\angle S_k(\omega)} + |N_k(\omega)|e^{i\angle N_{1,k}(\omega)} \quad (36)$$

$$X_{2,k}(\omega) = |S_k(\omega)|e^{i\angle S_k(\omega)-i\omega\tau} + |N_k(\omega)|e^{i\angle N_{2,k}(\omega)}. \quad (37)$$

The phase error for this block is

$$\theta_k(\omega) = \angle X_{1,k}(\omega) - \angle X_{2,k}(\omega) - \omega\tau. \quad (38)$$

Using the relation $R_k(\omega) = |S_k(\omega)|^2/|N_k(\omega)|^2$, from (36) and (37) we have

$$\angle X_{1,k}(\omega) = \angle\left(\sqrt{R_k(\omega)}e^{i\angle S_k(\omega)} + e^{i\angle N_{1,k}(\omega)}\right) \quad (39)$$

$$\angle X_{2,k}(\omega) = \angle\left(\sqrt{R_k(\omega)}e^{i\angle S_k(\omega)-i\omega\tau} + e^{i\angle N_{2,k}(\omega)}\right). \quad (40)$$

Equation (38) can be reduced to

$$\theta_k(\omega) = \angle\left(\sqrt{R_k(\omega)} + e^{iy_1}\right) - \angle\left(\sqrt{R_k(\omega)} + e^{iy_2}\right) \quad (41)$$

where $y_1 = \angle N_{1,k}(\omega) - \angle S_k(\omega)$ and $y_2 = \angle N_{2,k}(\omega) + \omega\tau - \angle S_k(\omega)$.

Usually, it is difficult to know the exact value of $y_1$ and $y_2$ since that requires exact knowledge about the signal and noise phases. However, even without such information, it is easy to show that, for $R_k(\omega) \geq 1$

$$\left|\angle\left(\sqrt{R_k(\omega)} + e^{iy_1}\right)\right| \leq \arcsin\left(\frac{1}{\sqrt{R_k(\omega)}}\right) \quad (42)$$

which, using (41), becomes

$$\left|\theta_k(\omega) + \angle\left(\sqrt{R_k(\omega)} + e^{iy_2}\right)\right| \leq \arcsin\left(\frac{1}{\sqrt{R_k(\omega)}}\right). \quad (43)$$

Now, as in (42), we have

$$\left|\angle\left(\sqrt{R_k(\omega)} + e^{iy_2}\right)\right| \leq \arcsin\left(\frac{1}{\sqrt{R_k(\omega)}}\right) \quad (44)$$

which, when combined with (43), becomes

$$\left|\frac{\theta_k(\omega)}{2}\right| \leq \arcsin\left(\frac{1}{\sqrt{R_k(\omega)}}\right). \quad (45)$$

Rearranging (45), we obtain

$$R_k(\omega) \leq \frac{1}{\sin^2\left(\frac{\theta_k(\omega)}{2}\right)}. \quad (46)$$

## References

[1] Voice Extreme Module Documentation [Online]. Available: http://www.sensoryinc.com/html/support/docs/80–0165-O.pdf

[2] P. Aarabi, "Genetic sensor selection enhanced independent component analysis and its applications to speech recognition," in *Proc. 5th IEEE Workshop Nonlinear Signal Information Processing*, June 2001.

[3] P. Aarabi and N. H. Khameneh, "Robust speech separation using visually constructed speech signals," in *Proc. 6th SPIE Conf. Sensor Fusion*, Apr. 2002.

[4] P. Aarabi and G. Shi, "Multi-channel time-frequency data fusion," in *Proc. 5th Int. Conf. Information Fusion*, Aug. 2002.

[5] P. Aarabi and S. Zaky, "Iterative spatial probability based sound localization," in *Proc. 4th World Multiconference Circuits, Systems, Computers, Communications*, July 2000.

[6] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[7] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, July 1995.

[8] F. Berthommier and S. Choi, "Comparative evaluation of casa and bss models for subband cocktail party speech segregation," in *Proc. Int. Conf. Spoken Language Processing*, Sept. 2002.

[9] J. Bitzer, K. U. Simmmer, and K. D. Kammeyer, "Multi-microphone noise reduction techniques for hands-free speech recognition-a comparative study," in *Proc. ROBUST*, May 1999, pp. 171–174.

[10] R. L. Bouquin-Jeanes, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 484–487, Sept. 1997.

[11] M. S. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Comput., Speech, Lang.*, vol. 9, pp. 153–169, 1995.

[12] M. S. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Conf. Acoustics, Speech, Signal Processing*, May 1997.

[13] K. Guentchev and J. Weng, "Learning-based three dimensional sound localization using a compact noncoplanar array of microphones," in *Proc. AAAI Symp. Intelligent Environments*, 1998.

[14] A. Hyvarinen, "Gaussian moments for noisy independent component analysis," *IEEE Signal Processing Letters*, vol. 6, pp. 145–147, June 1999.

[15] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[16] C. H. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-24, pp. 320–327, Aug. 1976.

[17] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 240–259, May 1998.

[18] I. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proc. 2001: A Speaker Odyssey*, June 2001.

[19] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[20] D. Rabinkin, "A DSP implementation of source location using microphone arrays," in *Proc. 131st Acoustical Society America*, May 1996.

[21] G. Shi, "Phase-Error Based Speech Enhancement," M.S. thesis, Dept. Elect. Comput. Eng., University of Toronto, Toronto, ON, Canada, 2002.

[22] E. Tessier, F. Berthommier, H. Glotin, and S. Choi, "A model of source segregation using the localization cue for robust cocktail-party speech recognition," in *Proc. Int. Conf. Speech Processing Elect. Comput. Eng.*, 1999.

[23] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Clliffs, NJ: Prentice-Hall, 1985.

**Parham Aarabi** (S'97–M'01) received the B.A.Sc. degree in engineering science from the University of Toronto, Toronto, ON, Canada, in 1998, the M.A.Sc. degree in electrical and computer engineering from the University of Toronto, in 1999, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 2001.

He is a Canada Research Chair in multisensor information systems, an Assistant Professor in the Edward S. Rogers, Sr., Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada, and the Founder and Director of the Artificial Perception Laboratory, University of Toronto. His current research interests include multi-sensor information fusion, human–computer interactions, and VLSI implementation of sensor fusion algorithms.

Dr. Aarabi has been the recipient of numerous teaching and research awards, including the Ontario Distinguished Researcher Award, the 2002 Fall Session Best Computer Engineering Professor Award, and the 2002/2003 Faculty of Engineering Early Career Teaching Award.

**Guangji Shi** received the M.A.Sc. degree in electrical and computer engineering from the University of Toronto (UT), Toronto, ON, Canada, and is currently pursuing the Ph.D. degree in microphone array based phase-error filtering at the same university.

He has worked in the automation industries both as a Technical Engineer and as a Software Developer. His current research interests include robust speech recognition, microphone arrays, and probabilistic reasoning.