ROBUST DIGIT RECOGNITION USING PHASE-DEPENDENT TIME-FREQUENCY MASKING

Guangji Shi   *Parham Aarabi*

Department of Elec. and Comp. Engineering, University of Toronto, Toronto, Ontario, Canada
guangji@comm.utoronto.ca   parham@ecf.utoronto.ca

ABSTRACT

A technique using the time-frequency phase information of two microphones is proposed to estimate an ideal time-frequency mask using time-delay-of-arrival (TDOA) of the signal of interest. At a signal-to-noise ratio (SNR) of 0dB, the proposed technique using two microphones achieves a digit recognition rate (average over 5 speakers, each speaking 20-30 digits) of 71%. In contrast, delay-and-sum beamforming only achieves a 40% recognition rate with two microphones and 60% with four microphones. Superdirective beamforming achieves a 44% recognition rate with two microphones and 65% with four microphones.

1. 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 [1]. This corruption often degrades the performance of these systems, for example, in speech recognition systems noise results in a lower speech recognition accuracy rate [1,2]. As a result, various speech enhancement techniques have been investigated in the past [7-10].

In such applications, multi-microphone based speech enhancement techniques have shown better promise compared to single-microphone based techniques [7,8]. Examples of multi-microphone techniques include Independent Component Analysis (ICA) [9,10] and various beamforming algorithms [3,4,7,8]. Beamforming has been extensively employed because of its robustness and simplicity [3,4].

Neither ICA nor beamforming, however, take advantage of the specific characteristics of speech signals. In [5], a time-frequency masking strategy that utilized only the phase information of the signals was proposed. It was shown that such a technique can be specifically useful for speech, since while a mixed speech signal (with one speaker of interest and several noise speakers) may be inseparable in either the time or frequency domain, they are in certain cases separable (to a certain extent) in the time-frequency domain. In fact, using the time-frequency technique of [5], average SNR gain of up to 15dB was obtained on noisy speech signals recorded by two microphones.

In this paper, the method of [5] is presented more thoroughly and the results are directly compared to two alternative techniques: delay-and-sum beamforming [3] and superdirective beamforming [7]. While the proposed technique requires knowledge regarding the TDOA of the speech signal of interest, this requirement is no different from that of other techniques, such as the beamforming techniques.

In section 2, we briefly review the delay-and-sum beamformer and superdirective beamformer. In section 3, the basic time-frequency masking ideas are presented. In sections 4, we compare the performance of the proposed technique with those of beamforming techniques in the context of a digit recognition system.

2. PROBLEM STATEMENT AND PRIOR WORK

Given M microphones in an environment, we model the signals received by the microphones as [6]:

$$\mathbf{x}(t) = \mathbf{h}(t) * s(t) + \mathbf{n}(t) \qquad (1)$$

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, equation (1) can be restated in the frequency domain as:

$$\mathbf{X}(\omega) = \mathbf{H}(\omega)S(\omega) + \mathbf{N}(\omega) \qquad (2)$$

where the capital letters are all Fourier Transforms of their lower-cased time domain representations. Note that because we are taking Fourier Transforms of finite signal segments, our representation in the frequency domain is a discrete one (i.e. ω can take on 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 equation (1), 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:

$$S(\omega) = \mathbf{A}(\omega)\mathbf{X}(\omega) \qquad (3)$$

where $\mathbf{A}(\omega)$ is a row of complex weights defined as follows [7]:

$$\mathbf{A}(\omega) = \frac{\mathbf{\Gamma}^{-1}(\omega)\mathbf{d}(\omega)}{\mathbf{d}^{*}(\omega)\mathbf{\Gamma}^{-1}(\omega)\mathbf{d}(\omega)} \qquad (4)$$

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

$$\mathbf{d}(\omega) = [1, e^{-j\omega\tau_2}, \cdots, e^{-j\omega\tau_M}] \qquad (5)$$

where $\tau_2, \tau_3, \ldots, \tau_M$ are the set of TDOAs for the 2nd 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

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & \Gamma_{X_1X_2} & \cdots & \Gamma_{X_1X_M} \\ \Gamma_{X_2X_1} & 1 & \cdots & \Gamma_{X_2X_{M-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{X_MX_1} & \Gamma_{X_MX_3} & \cdots & 1 \end{pmatrix} \qquad (6)$$

For delay-and-sum beamforming, we have [3,7]:

$$\Gamma_{X_uX_v}(\omega) = 0 \text{, for } u \neq v \qquad (7)$$

For superdirective beamforming, we have [7]:

$$\Gamma_{X_uX_v}(\omega) = \frac{\sin\left(\frac{\omega d_{uv}}{c}\right)}{\frac{\omega d_{uv}}{c}\left(1 + \frac{\sigma_n^2}{P_{NN}(\omega)}\right)} \qquad (8)$$

where *c* is the speed of sound, d_{uv} is distance between the *u*th and *v*th microphones, σ_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 [7], a -20dB to -40dB sensor noise to room noise ratio gives good results in most practical situations.

3. PHASE-DEPENDENT TIME-FREQUENCY MASKING FUNCTION

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 (the windowing function is chosen so 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 ω takes on a set of discrete frequencies (in steps of $2\pi F_s / N$) values due to the finite time-window. $X_k(\omega)$ 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 IFFT of $X_k(\omega)$ for each segment, overlapping and adding the segments, and reconstructing the continuous signal from the discrete-time signal.

Because certain time-frequency blocks are more dominant for speech signals [1], and in the case of multiple speakers, these can often be different blocks, our goal is to find a time-frequency masking function $H_k(\omega)$ such that:

$$Y_k(\omega) = X_k(\omega)H_k(\omega) \qquad (9)$$

results in the signal $y(t)$ (which is the inverse TF transform of $Y_k(\omega)$) has weaker signal components from undesired speech sources. In effect, $H_k(\omega)$ behaves like a reward-punish algorithm. TF components with small SNRs would have a TF mask with a high value, and noisy TF blocks would have a TF mask close to zero. It can be shown that if the SNR of a given TF, defined by:

$$R_k(\omega) = \frac{|S_k(\omega)|^2}{|N_k(\omega)|^2} \qquad (10)$$

is known, then the ideal (SNR-maximizing) TF-mask would be [2]:

$$H_k^{*}(\omega) = \frac{R_k(\omega)}{1 + R_k(\omega)} \qquad (11)$$

While the actual TF-block SNR $R_k(\omega)$ is difficult to estimate in general, it becomes possible to obtain an upper bound for it in the case where we have two microphones. Assuming the two microphones receive the signals $x_1(t)$ and $x_2(t)$, respectively, we model the two signals as follows:

$$\begin{aligned} x_1(t) &= s(t) + n_1(t) \\ x_2(t) &= s(t-\tau) + n_2(t) \end{aligned} \tag{12}$$

where $s(t)$ is the signal corresponding to the speech source of interest arriving at the first microphone, $n_1(t)$ and $n_2(t)$ are possibly dependent noise signals, and τ is a known TDOA corresponding to the speech source of interest. Note that while we have not included reverberation in the above model, the reverberation can be included as part of the noises since no assumption about the independence of the noises and the speech signal is made. In the TF-domain, we thus have:

$$\begin{aligned} X_{1,k}(\omega) &= S_k(\omega) + N_{1,k}(\omega) \\ X_{2,k}(\omega) &= S_k(\omega)e^{-j\omega\tau} + N_{2,k}(\omega) \end{aligned} \tag{13}$$

If we assume that the noise magnitudes in the TF-domain are equal for both microphones (i.e. $|N_{1,k}(\omega)|=|N_{2,k}(\omega)|$), which is a valid assumption as long as the inter-microphone distance is small, then we obtain [2]:

$$R_k(\omega) \leq \frac{1}{\sin^2\left(\frac{\theta_k(\omega)}{2}\right)} \tag{14}$$

where the phase-error $\theta_k(\omega)$ is defined as:

$$\theta_k(\omega) = \angle X_{1,k}(\omega) - \angle X_{2,k}(\omega) - \omega\tau \tag{15}$$

Note that the phase-error above is always wrapped to the range $[-\pi, \pi]$ in this paper. This implies that our ideal TF mask has the following upper bound:

$$H_k^*(\omega) \leq \frac{1}{1+\sin^2\left(\frac{\theta_k(\omega)}{2}\right)} \tag{16}$$

Hence, the phase-error $\theta_k(\omega)$ can be used to define the upper bound on the reward-punish factor for each TF block. The above mask bound motivates the investigation of various phase-error dependent masks, the following one of which will be explored in this paper:

$$H_k(\omega) = \frac{1}{1+\gamma\theta_k^2(\omega)} \tag{17}$$

where γ is a constant defining the aggressiveness of the mask (higher γ values correspond to more aggressive masks and vice versa). Depending on the value of γ, the phase-error mask punishes each and every frequency block based on its phase-error. Blocks with large phase-errors are punished more and blocks with small phase-errors are punished less.

The TF masking function obtained from the phases of the two microphone signals (in the TF-domain) can be applied to each of the two signals separately. Throughout this paper, when we refer to the output of the TF-masking procedure, we imply the application of the mask to one of the two signals.

The effectiveness of the proposed mask is explored by a simulation consisting of two female speakers (one being considered as noise). The main speaker and the noise speaker had a 3 sample and –5 sample delay, respectively. The volume of the noise speaker was adjusted to result in the desired input SNR. Each input speech segment contains 200,000 samples sampled at 16 KHz. The large speech segment is decomposed into 400-sample segments using a Hanning window. The resulting noisy channels are processed by the proposed TF masking procedure, the results of which are shown in Figure 3. When the input SNR is less than 20dB, a significant SNR gain is obtained. Furthermore, aggressive phase-error masking (high γ) is good for low SNR input while less aggressive phase-error masking is better for high SNR input (low γ).

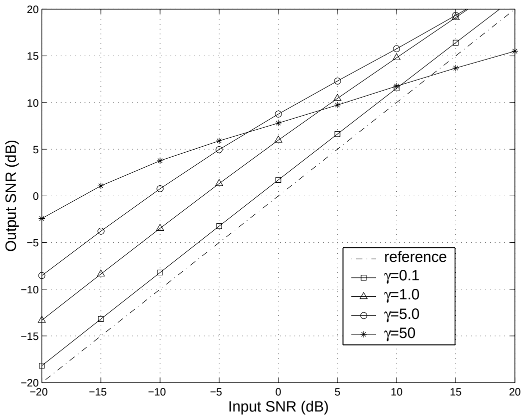

Figure 3: Effect of the proposed TF mask on SNR gains.

4. PERFORMANCE EVALUATIONS

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 5 different speakers. A speaker-independent single digit recognition system (with no training) was built based on the Voice Extreme Module from Sensory Inc. The speech recognition system, which is small enough to be embedded in handheld applications, recorded 20-30 random digits (in the 0-9 range) from each speaker. The noise was artificially-added speech noise consisting of a male conversation. The source of interest and the noise source had a 10 sample and 5 sample delay, respectively. The sampling frequency was 16 KHz. For the phase-error TF mask, equation (17) was used with γ=5.

The experimental results are shown in Tables 1 and 2. The results show that the phase-dependent TF masking has approximately a 20% higher recognition accuracy rate than the other two techniques. Notice the masking technique is able to achieve a recognition rate at 0dB SNR that is slightly better than those of beamforming techniques at 10dB SNR. The digit recognition results agree with the obtained SNR gains. The average SNR gain generated by delay-and-sum beamforming and superdirective beamforming was 1.3dB and 2.4dB, respectively. On the other hand, the TF masking technique was able to generate approximately 8dB gain when the input SNR is 0dB and 5dB gain when the input SNR is 10dB.

Table 1: Digit recognition rate comparison (DS=delay-and-sum beamforming, SD=superdirective beamforming, TF=time frequency masking) with 2 microphones at 0dB.

	S 1	**S 2**	**S3**	**S4**	**S 5**	**Average**
DS	36%	36%	64%	57%	10%	**40%**
SD	41%	36%	64%	62%	15%	**44%**
TF	59%	77%	86%	71%	60%	**71%**

Table 2: Digit recognition rate comparison (2 microphones, 10dB SNR)

	S1	**S2**	**S3**	**S4**	**S5**	**Average**
DS	68%	64%	77%	76%	35%	**64%**
SD	73%	64%	77%	71%	35%	**64%**
TF	77%	82%	86%	81%	80%	**81%**

To further compare these three techniques, another experiment was conducted to evaluate the performance of the two beamforming techniques using 4 microphones with 0dB SNR. In this case, the average SNR gain generated by delay-and-sum beamforming and superdirective beamforming was 4.4dB and 8.5dB, respectively. Table 3 shows the recognition rate for this case. Although the SNR gain achieved by superdirective beamforming is almost twice that of delay-and-sum beamforming, no significant digit recognition rate was observed. From Tables 1 and 3, it is evident that the performance of the proposed TF masking technique with 2 microphones is similar to the beamforming techniques with 4 microphones.

Table 3: Digit recognition rate comparison (4 microphones, 0dB SNR)

	S1	**S2**	**S3**	**S4**	**S5**	**Average**
DS	64%	59%	73%	76%	30%	**60%**
SD	73%	73%	77%	71%	30%	**65%**

5. CONCLUSIONS

A phase-error dependent time-frequency masking technique was proposed. It was shown through speaker independent digit recognition experiments that the proposed technique achieves a substantially higher recognition rate than prior superdirective and delay-and-sum beamforming techniques. It should be noted that the proposed time-frequency masking technique operates on each channel separately. As a result, other multi-channel algorithms can be used to result in further improvement.

6. REFERENCES

[1] L. Rabiner and B. Juang. Fundamentals of speech recognition, Prentice-Hall, New Jersey, 1993.

[2] G. Shi. Phase-error based speech enhancement, M.A.Sc. Thesis, Department of Electrical and Computer Engineering, University of Toronto, 2002.

[3] D.H. Jonhson, D.E. Dungeon. Array Signal Processing: Concepts and Techniques, Prentive-Hall, 1993.

[4] B. Widrow and S. D. Stearns. Adaptive signal processing, Prentice-Hall, New Jersey, 1985.

[5] P. Aarabi and G. Shi. “Multi-channel time-frequency data fusion,” In Proceedings of 5th International Conference on Information Fusion, Washington D.C., July 2002.

[6] M.S. Brandstein and H. Silverman. “A robust method for speech signal time-delay estimation in reverberant rooms,” Proceedings of ICASSP, May 1996.

[7] J. Bitzer, K.U. Simmmer and K.D. Kammeyer, “Multi-microphone noise reduction techniques for hands-free speech recognition-a comparative study,” Proceeding of ROBUST’99, pp. 171-174, Tampere, Finland, May 1999.

[8] I. McCowan, J. Pelecanos, and S. Sridharan, "Robust Speaker Recognition using Microphone Arrays," in Proceedings of 2001: A speaker odyssey, June 2001.

[9] A. Hyvrinen and E. Oja. Independent component analysis: Algorithms and applications. Neural networks, 13(4-5):411:430, 2000

[10] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. Neural Comp., 7:1129-1159, July 1995