MULTIPLE-MICROPHONE TIME-VARYING FILTERS FOR ROBUST SPEECH RECOGNITION

Calvin Yiu-Kit Lai *Parham Aarabi*

The Artificial Perception Lab
The Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto
calvin.lai@utoronto.ca parham@ecf.utoronto.ca

ABSTRACT

A multiple microphone time varying filter that is an extension of the dual-microphone speech enhancement technique of [7] is proposed and experimentally analyzed. The technique utilizes information regarding the locations of the speech source of interest and the microphones to compute a time varying filter that results in substantial noise reduction over other speech enhancement techniques such as delay-and-sum beamforming and superdirective beamforming. For example, digit recognition results in an environment with two speakers and a reverberation time of 0.1s show a recognition accuracy rate increase of 25.2% over delay-and-sum beamforming and an increase of 26.5% over superdirective beamforming using six microphones.

1. INTRODUCTION

The recognition of speech by computers will one day transform how and where computers are used in our society. Before this can happen, however, automatic speech recognition systems must become robust to noise and reverberation, and be able to withstand the confusion and difficulty that they currently encounter in the presence of secondary speech sources.

Clearly, single microphone techniques can only go so far [4, 1, 7]. Multiple microphones enable us to separate speech signals based on their spatial origin [7, 6]. Such a separation can result in a significant noise reduction, which then enables robust speech recognition to be performed. Achieving significant noise reductions is not easy. There have been numerous techniques proposed through the years, including various beamforming techniques [2, 6], Independent Component Analysis (ICA) [6, 9], as well as numerous other techniques. One of the more successful of these techniques is the dual-microphone approach of [7] and [6]. This technique, while only valid for two-microphones, was shown to perform much better than techniques such as superdirective and delay-and-sum beamforming.

The prior work on this technique, which includes [7, 8, 6], has left two questions open. The first question regards the validity and effectiveness of the algorithm in a realistic environment (i.e. an actual experiment in a reverberant and noisy room). The second question regards the benefit and possibility of extending this technique to multiple microphones. Both of these questions are addressed in this paper.

2. PROBLEM STATEMENT AND PRIOR WORK

We assume that an array of M microphones is available to record a speech source in a noisy and reverberant environment. The signal observed by the ith microphone can be modeled as:

$$x_i(t) = h_i(t) * s(t) + n_i(t) \tag{1}$$

where $h_i(t)$ is the impulse response corresponding to the speech source and the ith microphone, $s(t)$ is the original signal produced by the speech source, and $n_i(t)$ is the noise component of the ith microphone. Now, given these recordings, our goal is to somehow obtain the original signal $s(t)$ without any knowledge of either the impulse responses or the noises. In this paper, we will assume that the location of the microphones *and* the location of the speech source of interest are known. The problem is illustrated in Figure 1.

2.1. Preliminaries

In practice, it is easier to consider the frequency domain version of equation 1, which can be stated as:

$$X_i(\omega) = H_i(\omega)S(\omega) + N_i(\omega) \tag{2}$$

where capital letters represent the Fourier transforms of their lower-cased time-domain signals. In practice, since our Fourier transform (which in reality is a Fast Fourier Transform or FFT) is perform over a discrete and finite segment (say, with a total of N samples), our frequency representation

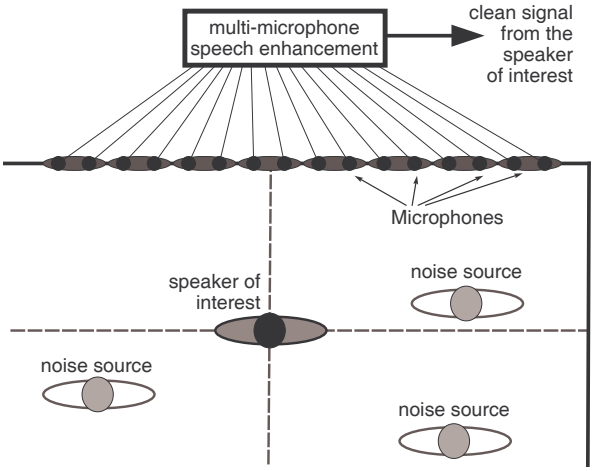

Fig. 1. The multi-microphone speech enhancement problem

will be discrete one, starting from a frequency of $-\pi F_s/2$ upto $\pi F_s/2$ in steps of $2\pi F_s/N$, where F_s is the sampling rate. In order to improve the smoothing effect of the finite segment size, each time segment is multiplied by a Hanning (which more correctly, should be called Von Hann) window. Furthermore, consecutive segments are half-overlapped such that after processing, the resulting segments can be half-overlapped and added to reconstruct the desired signal.

2.2. Time Varying Phase-Based Dual-Microphone Filters

One of the most successful solutions to the previously stated problem is time varying phase-based filtering techniques described by [8, 6, 7]. These techniques, which until now, could only be applied to the two channel case, utilize the phase difference between the signals of the different microphones as well as the expected time delays of arrivals (TDOAs) in order to estimate the signal-to-noise ratio SNR of *each* frequency component. The scaling or filtering of that frequency component will then be performed corresponding to its SNR estimate. Assuming that the TDOA corresponding to microphone i and the speaker location is τ_i (this can be easily estimated from prior knowledge regarding the microphone and speaker locations), the time-varying phase-based filter is defined as:

$$H_{12}(\omega) = \rho_{12}(\omega)\eta_{12}(1,\omega) + (1-\rho_{12}(\omega))\eta_{12}(-1,\omega) \quad (3)$$

where $\eta_{12}(\mu,\omega)$ is defined as:

$$\eta_{12}(\mu,\omega) = \frac{\frac{|X_1(\omega)|}{2|X_2(\omega)|} + \frac{|X_2(\omega)|}{2|X_1(\omega)|} + \mu}{\frac{|X_1(\omega)|}{|X_2(\omega)|} + \frac{|X_2(\omega)|}{|X_1(\omega)|} + \mu - \cos(\theta_{12}(\omega))} \quad (4)$$

where the phase error $\theta_{12}(\omega)$ is defined as:

$$\theta_{12}(\omega) = \angle X_1(\omega) - \angle X_2(\omega) - \omega(\tau_1 - \tau_2) \quad (5)$$

and $\rho_{12}(\omega)$ is defined as:

$$\rho_{12}(\omega) = e^{-10(\theta_{12}(\omega))^2} \quad (6)$$

Note that the phase error must be wrapped between $-\pi$ and π in the above equation.

2.3. Beamforming

Another common technique that is used for speech enhancement using prior knowledge regarding the TDOAs is beamforming [1, 2, 6]. Here, we will consider two beamforming techniques: delay-and-sum beamforming (DSB) and superdirective beamforming (SDB)[1, 2].

The DSB output can be stated as follows:

$$Y_{\mathrm{DSB}}(\omega) = \frac{\mathbf{d}^H(\omega)\mathbf{X}(\omega)}{\mathbf{d}^H(\omega)\mathbf{d}(\omega)} \quad (7)$$

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

$$\mathbf{d}(\omega) = \left[\begin{array}{cccc} e^{-jw\tau_1} & e^{-jw\tau_2} & \dots & e^{-jw\tau_M} \end{array}\right]^T \quad (8)$$

and the frequency data vector $\mathbf{X}(\omega)$ is defined as:

$$\mathbf{X}(\omega) = \left[\begin{array}{cccc} X_1(\omega) & X_2(\omega) & \dots & X_M(\omega) \end{array}\right]^T \quad (9)$$

A more successful beamforming technique, known as superdirective beamforming, has been widely used for speech enhancement in practical and realistic conditions [1, 3].

In SDB, the signal received by each microphone is filtered and summed across all microphones, resulting in the following output:

$$Y_{\mathrm{SDB}}(\omega) = \frac{1}{M}\mathbf{W}^H(\omega)\mathbf{X}(\omega) \quad (10)$$

where the weight vector is defined as $\mathbf{W}(\omega) = \left[\begin{array}{ccc} W_1(\omega) & W_2(\omega) & \dots \end{array}\right.$ and can be obtained using the coherence matrix as follows:

$$\mathbf{W}(\omega) = \frac{{\mathbf{\Gamma_{VV}}}^{-1}(\omega)\mathbf{d}(\omega)}{\mathbf{d}^H(\omega){\mathbf{\Gamma_{VV}}}^{-1}(\omega)\mathbf{d}(\omega)} \quad (11)$$

The coherence matrix is defined as:

$$\mathbf{\Gamma_{VV}}(\omega) = \begin{pmatrix} 1 & \Gamma_{V_1V_2}(\omega) & \dots & \Gamma_{V_1V_M}(\omega) \\ \Gamma_{V_2V_1}(\omega) & 1 & \dots & \Gamma_{V_2V_M}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{V_MV_1}(\omega) & \Gamma_{V_MV_2}(\omega) & \dots & 1 \end{pmatrix} \quad (12)$$

using:

$$\Gamma_{V_iV_j}(\omega) = \mathrm{sinc}\left(\frac{\omega d_{ij}}{c}\right) \quad (13)$$

where d_{ij} is the distance between the ith and jth microphones and c is the speed of sound in air (about 345m/s).

3. THE EXTENSION OF PHASE BASED TIME VARYING FILTERS TO M MICROPHONES

While DSB and SDB can be readily applied to a situation with any number of microphones, the phase-based time varying filter defined previously can only be applied to two microphones. In this paper, we have a total of M microphones available. This means that for each microphone, a total of $M-1$ microphone pairs can be formed. Hence, for each microphone, a total of $M-1$ time-varying filters can be obtained.

After a detailed initial analysis (which can be found in [5]), it was discovered that a modified geometric mean of the time-varying filters would provide the best results. In other words, our overall filter $\Psi(\omega)$ for the ith microphone can be defined as:

$$\Psi_i(\omega) = \left(\prod_{j=1, j\neq i}^{M} (H_{ij}(\omega)) \right)^{1/k} \tag{14}$$

where $H_{ij}(\omega)$ is the time varying phase-based filter obtained from the ith and jth microphone pair combination and k is a value which for a standard geometric mean would be equal to M. In this case, it was experimentally discovered that a value of $k = 0.3M$ results in a more aggressive filtering strategy but in significantly improved separation results. If $k << 0.3M$, the overall filter becomes too aggressive resulting in the significant loss of both signal and noise. If $k >> 0.3M$, then a significant portion of the noise remains which still prevents the proper recognition of the output.

After the geometric mean of equation 14 is used to obtain the overall filter for *each* microphone, the filter is applied to the signal of the microphone and then delay-and-sum beamforming is performed on the M filtered signals, as shown below:

$$Y(\omega) = \sum_{i=1}^{M} \Psi_i(\omega) X_i(\omega) e^{j\omega\tau_i} \tag{15}$$

4. EXPERIMENTAL EVALUATION

In order to evaluate the performance of the proposed multi-microphone speech enhancement technique, which is the extension of the two-channel time varying phase-based noise removal technique of [7], a series of experiments were performed. Figure 2 illustrates the setup of the experiments. Each experiment involved two speakers, one speaking digits in the range 0-9, and the other speaking random words. The two speakers were synchronized to speak the digit/random word combination simultaneously (hence, making separation a more difficult). A total of 30 digits were spoken in each trial, and a total of 8 trials were conducted to ensure the validity of the results (resulting in a total of 240 digit recognitions). Only the first six microphones (the six on the left of Figure 2) were used.

The overall signal-to-noise ratio between the two speakers was 0dB. The background noise resulted in a sensor signal-to-noise ratio of approximately 20dB. The reverberation time of the setup was approximately 0.1s.

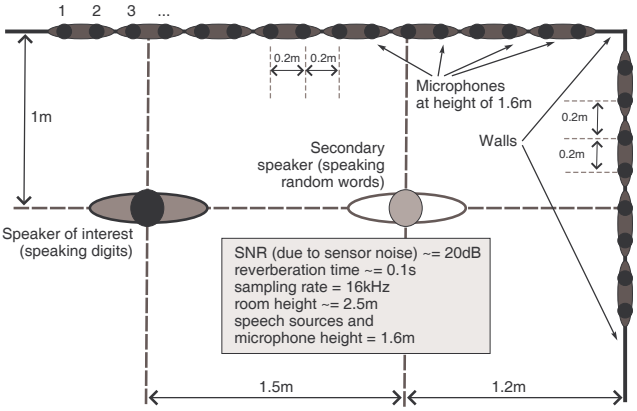

Fig. 2. Room configuration showing 16 microphones on a wall with a fixed inter-microphone distance of 0.2m.

A digit recognition module (Sensory Inc.'s Voice Extreme) was used for digit recognition experiments for the desired speaker (the one speaking digits only). The module is a speaker-independent digit recognition system that can achieve about an 85% recognition accuracy rate in an ideal and noiseless environment.

The signals that were recorded from the microphones were used as the basis of DSB, SDB, and the technique proposed in this paper. They were segmented into half-overlapped 1024-sample time segments, processed according to one of the algorithms, and then overlapped and added to reconstruct the desired speech signal. The digit recognition rate comparison using different numbers of microphones and different algorithms is shown in Figure 3.

As shown, the technique proposed in this paper far outperforms the other two technique as well as the mixed signal. For the six microphone case, the recognition rate of the Multi-Channel Time-Varying Phase-based (MC-TVP) technique proposed here is 31.9% greater than that of the mixed signal, 25.2% greater than that of DSB, and 26.5% greater than that of SDB. It should be emphasized that these results were obtained using 240 separate speaker independent digit recognitions in a *real* reveberant environment. This is partially why the performance of SDB is similar to that of DSB.

5. SUMMARY

In this paper, a multi-microphone phase-based time varying filter was proposed. The basic two-channel version of this

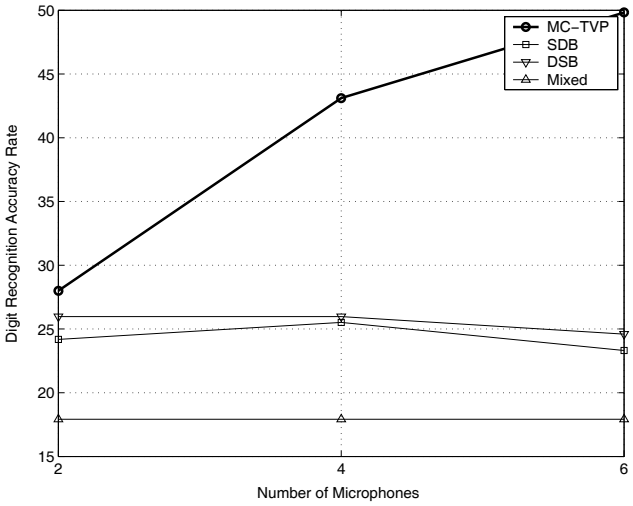

Fig. 3. Digit recognition rate comparison between multi-channel time varying phase-based filters (MC-TVP), delay-and-sum beamforming (DSB), superdirective beamforming (SDB), and an unprocessed microphone signal (which would be the 6^{th} microphone).

filter was previously presented in [7]. However, the work here extends this to the M microphone case *and* evaluates its performance in a real environment.

In terms of the recognition rate, it is clear that the proposed technique outperforms DSB and SDB specially when there are more microphones available. With only two microphones, the difference between the three techniques is not significant. One thing that does not appear in the recognition results is the quality of the audio that results after the application of the filter proposed in this paper. The perceptual quality of the separated signal (using the technique proposed in this paper) is substantially better than that of either DSB or SDB, for 2, 4, *and* 6 microphones.

Clearly, the experimental analysis performed is limited in that it is based upon a speaker independent digit recognition system. In order to truly evaluate the value of the proposed algorithm, a detailed study must be done using current state-of-the-art speaker dependent speech recognition systems. Without such experiments, this paper and the results therein are only a point of validation of an algorithm that requires further investigation.

Finally, the comparisons in this paper are only meant as a mechanism of benchmarking the proposed algorithm. Numerous techniques, such as postfiltering (which, by the way, is also a time varying filter) can achieve considerable speech separation results. In this paper, postfiltering was not compared but clearly, this is an avenue of future research and emphasis [5].

6. REFERENCES

[1] J. Bitzer, K. Uwe Simmer, Superdirective Microphone Arrays, in Microphone Arays, Cambridge, MA, January 2001, pg. 19-38.

[2] J. Bitzer, K.U. Simmmer and K.D. Kammeyer. Multimicrophone noise reduction techniques for hands-free speech recognition-a comparative study. Proceeding of Robust Methods for Speech Recognition in Adverse Conditions (ROBUST-99), pp. 171-174, Tampere, Finland, May 1999.

[3] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.35 pp. 1365-1375, Oct 1987.

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

[5] C.Y.K. Lai, Analysis and Extension of Time-Frequency Masking, M.A.Sc. Thesis, Department of Electrical and Computer Engineering, University of Toronto, 2003.

[6] G. Shi, P. Aarabi, Robust Digit Recognition Using Phase-Dependent Time-Frequency Masking. Proceedings of the 2003 IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, April 2003.

[7] P. Aarabi, G. Shi, O. Jahromi, Robust Speech Separation Using Time-Frequency Masking. Proceedings of the 2003 IEEE Conference on Multimedia and Expo (ICME 2003), Baltimore, Maryland, July 2003.

[8] 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.

[9] P. Aarabi, Genetic sensor selection enhanced independent component analysis and its applications to speech recognition. In Proceedings of the 5th IEEE Workshop on Nonlinear Signal and Information Processing, June 2001.

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