MULTI-MICROPHONE CORRELATION-BASED PROCESSING FOR ROBUST SPEECH RECOGNITION

Thomas M. Sullivan and Richard M. Stern

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

In this paper we present a new method of signal processing for robust speech recognition using multiple microphones. The method, loosely based on the human binaural hearing system, consists of passing the speech signals detected by multiple microphones through bandpass filtering and nonlinear rectification operations, and then cross-correlating the outputs from each channel within each frequency band. These operations provide an estimate of the energy contained in the speech signal in each frequency band, and provides rejection of off-axis jamming noise sources. We demonstrate that this method increases recognition accuracy for a multi-channel signal compared to equivalent processing of a monaural signal.

1. INTRODUCTION

The need for speech recognition systems and spoken language systems to be robust with respect to their acoustical environment has become more widely appreciated in recent years. Results of several studies have demonstrated that even automatic speech recognition systems that are designed to be speaker independent can perform very poorly when they are tested using a different type of microphone or acoustical environment from the one with which they were trained, even in a relatively quiet office environment (*e.g.* [1]). Applications such as speech recognition over telephones, in automobiles, on a factory floor, or outdoors demand an even greater degree of environmental robustness. The proposed paper describes a novel algorithm for combining the outputs of multiple microphones that improves the recognition accuracy of automatic speech recognition systems.

Several different types of array processing strategies have been applied to speech recognition systems. The simplest such system is the delay-and-sum beamformer (*e.g.* [2]). In delay-and-sum systems, steering delays are applied at the outputs of the microphones to compensate for arrival time differences between microphones to a desired signal, reinforcing the desired signal over other signals present. A second approach is to use an adaptive algorithm based on minimizing mean square energy, such as the Frost or the Griffiths-Jim algorithm [3]. These algorithms can provide nulls in the direction of undesired noise sources, as well as greater sensitivity in the direction of the desired signal, but they assume that the desired signal is statistically independent of all sources of degradation. Consequently, they do not perform well in environments when the distortion is at least in part a delayed version of the desired speech signal as is the case in many typical reverberant rooms (*e.g.* [4]). (This problem can be avoided by only adapting during non-speech segments [5]).

The algorithm described in this paper is based on a third type of processing, which is loosely motivated by the cross-correlation-based processing in the human binaural system. The human auditory system is a remarkably robust recognition system for speech in a wide range of environmental conditions, and other signal processing schemes have been proposed that are based on human binaural hearing (*e.g.* [6]). Nevertheless, most previous studies have used cross-correlation-based processing to identify the direction of a desired sound source, rather than to improve the quality of input for speech recognition (*e.g.* [7,8]).

We describe the new cross-correlation-based algorithm in the following section. We describe the ability of the algorithm to preserve the shape of vowel spectra in Section 3, and in Section 4 we report on the results of pilot experiments in which the algorithm was used to improve speech recognition accuracy.

2. CROSS-CORRELATION-BASED MULTI-MICROPHONE PROCESSING

Figure 1 is a simplified block diagram of the multi-microphone correlation-based processing system. The input signals $x_k[n]$ are first delayed in order to compensate for differences in the acoustical path length of the desired speech signal to each microphone. The signals from each microphone are passed through a bank of bandpass filters with different center frequencies, passed through nonlinear rectifiers, and the outputs of the rectifiers at each frequency are correlated. Currently we use the 40-channel filterbank proposed by Seneff [9], which was designed to approximate the frequency selectivity of the auditory system. The shape of the rectifier has a significant effect on the results. We have examined the response of two types of nonlinear rectifiers: the rectifier originally described by Seneff, which saturates in its response to high-level stimuli, and a family of rectifiers called half-wave power-law rectifiers which produce zero output for negative signals and raise positive signals to an integer power.

For two microphones, these operations correspond to the familiar short-time cross-correlation operation for an arbitrary bandpass channel with center frequency ω_c:

$$E_c = \sum_{n=0}^{N-1} y_1[n, \omega_c]\, y_2[n, \omega_c]$$

where $y_k[n, \omega_c]$ is the signal from the k^{th} microphone after delay, bandpass filtering, and rectification, n is the time index, and N is the number of samples per analysis frame. For the general case of K microphones, this produces

$$\hat{E}_c = \{ \sum_{n=0}^{N-1} y_1[n, \omega_c] \prod_{k=2}^{K} y_k[n, \omega_c] \}^{2/K}$$

The factor of $2/K$ in the exponent enables the result to retain the dimension of energy, regardless of the number of microphones.

The 40 "energy" values are then converted into 40 cepstral coefficients using the cosine transform. The 40 cepstral parameters and an additional coefficient representing the power of the signal during the analysis frame are used as phonetic features for the original CMU SPHINX-I recognition system [10].

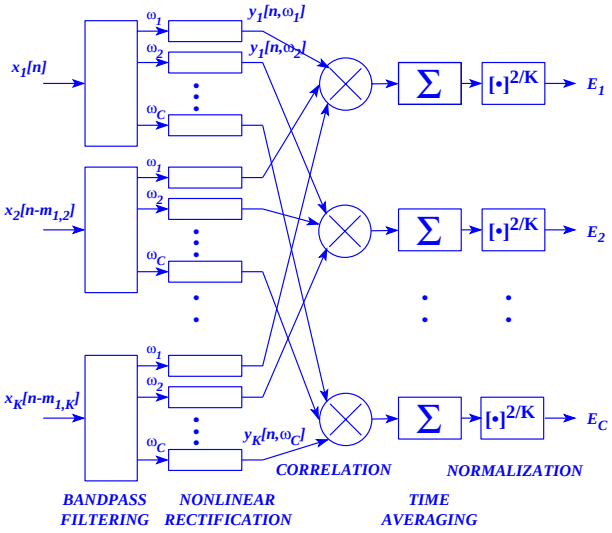
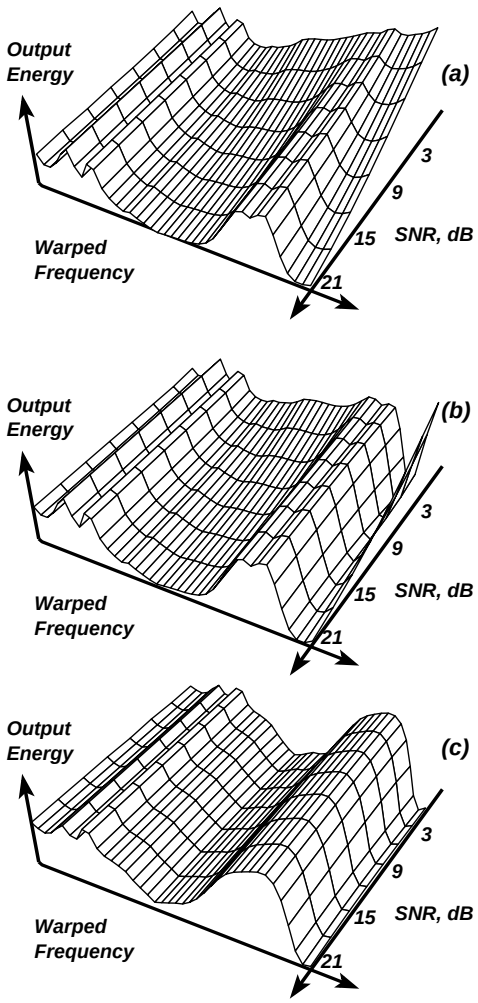


Figure 1. Block diagram of multi-microphone cross-correlation-based processing system.

3. EFFECTS OF CROSS-CORRELATION PROCESSING ON SPECTRAL PROFILES

We first confirmed the validity of the algorithm by an analysis of a digitized vowel segment /a/ corrupted by artificially-added white Gaussian noise at global SNRs of 0 to +21 dB. The speech segment was presented to all microphone channels identically (to simulate a desired signal arriving on axis) and the noise was presented with linearly increasing delays to the channels (to simulate an off-axis corrupting signal impinging on a linear microphone



igure 2. Estimates of spectra for the vowel segment /a/ for various SNR using (a) 2 input channels and zero delay, (b) 2 input hannels and 125-μs delay to successive channels, and (c) 8 input hannels and 125-μs delay.

array). We simulated the processing of such a system using 2 and 8 microphone channels, and time delays for the masking noise of 0 and 125 μs to successive channels.

Figure 2 describes the effect of SNR, the number of processing channels, and the delay of the noise on the spectral profiles of the vowel segment. The frequency representation for the vowel segment is shown along the horizontal axis. (These responses are warped in frequency according to the nonlinear spacing of the auditory filters.) The SNR was varied from 0 to +21 dB in 3-dB steps, as indicated. The upper panel summarizes the results that are obtained using 2 channels with the noise presented with zero delay from channel to channel (which would be the case if the speech and noise signals arrive from the same direction). Note that the shape of the vowel, which is clearly defined at high SNRs becomes almost indistinct at the lower SNRs. The center and lower panels show the results of processing with 2 and 8 microphones, respectively, when the noise is presented with a delay of 125 μs from

channel to channel (which corresponds to a moderately off-axis source location for typical microphone spacing). We note that as the number of channels increases from 2 to 8, the shape of the vowel segment in Figure 2 becomes much more invariant to the amount of noise present. In general, we found in our pilot experiments that the benefit to be expected from processing increases sharply as the number of microphone channels is increased. We also observed (unsurprisingly) that the degree of improvement increases as the simulated directional disparity between the desired speech signal and the masker increases. We conclude from these pilot experiments that the cross-correlation method described can provide very good robustness to off-axis additive noise. As the number of microphone channels increases, the system is robust to noise at smaller time delays between microphones, so even undesired signals that are slightly off-axis can be rejected.

4. EFFECTS OF CROSS-CORRELATION PROCESSING ON SPEECH RECOGNITION ACCURACY

Encouraged by the appearance of these spectral profiles with simulated input, we evaluated 2- and 4-channel implementations of the algorithm in the context of an actual speech recognition system. The CMU SPHINX-I speech recognizer [10] was trained using speech recorded in an office environment using the speaker-independent alphanumeric census database [1] with the omnidirectional desktop Crown PZM6FS microphone. Identical samples of 1018 training utterances from this database from 74 speakers were presented to the inputs of the multi-microphone system described in Figure 1. All speech was sampled at 16 kHz. The frame size for analysis was 20 ms (320 samples) and frames were analyzed every 10 ms. Two different testing databases were used, as described below.

4.1. Nonlinear Rectification

The goal of the first series of experiments using actual speech input to the system was to determine the effect of rectifier shape on speech recognition accuracy. A test database was collected using a stereo pair of PZM6FS microphones placed under the monitor of a NeXT workstation. The database consisted of 10 male speakers each uttering 14 alphanumeric census utterances that were similar to those in the training data.

We compared the word errors obtained (tabulated according to the standard DARPA metric) using a 2-channel implementation of the cross-correlation algorithm and a “mono” implementation of the same algorithm in which the same signal is input to the two channels. (The “mono” implementation enables us to assess the extent to which the system can exploit differences between the signals arriving at the two microphones.) We tested with half-wave power-law rectifiers with various exponents, and with the rectifier proposed by Seneff [9]. Figure 3 summarizes the results of these comparisons. Using the half-wave power-law rectifier with the positive signal raised to the 2nd power (the “half-square” rectifier) provided the lowest word error rate of the various half-wave

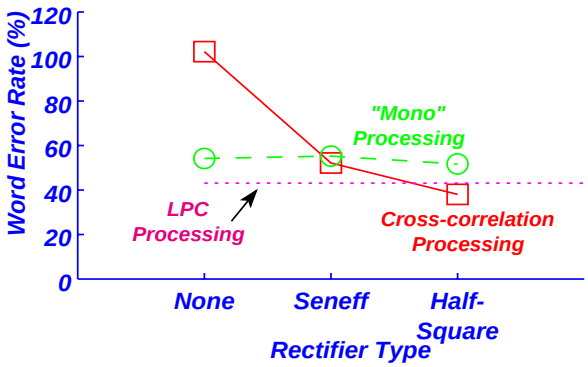


Figure 3. Comparison of word error rates achieved with 2-microphone processing using various half-wave rectifiers, and three types of signal processing.

power-law rectifiers. The results show that the 2-channel cross-correlation algorithm provides a slightly better error rate than conventional LPC signal processing, and that the recognition accuracy using this algorithm depends on the shape of the rectifier.

We hypothesize that the half-square rectifier provides the best error rate because it is slightly expansive. The Seneff rectifier actually compresses the positive signals and limits dynamic range. Using a power-law rectifier of too great a power starts to diminish in performance as the dynamic range is expanded too greatly. Using no rectifier at all provides poor performance because negative correlation values are produced. The half-wave square-law rectifier was used for all subsequent experiments.

4.2. Number of Processing Channels

We describe in this section initial results obtained using a new set of multiple-channel speech data. This testing database consisted of utterances from the census task, and was collected in a much more difficult environment with significant reverberation and additive noise sources. The ambient noise level was approximately 60 dB SPL with linear frequency weighting. Simultaneous speech samples from 10 new male speakers were collected using (1) a 4-element linear array of inexpensive noise-cancelling pressure gradient electret condenser microphones, spaced 7 cm from one another, (2) a pair of omnidirectional desktop Crown PZM6FS microphones, also spaced 7 cm from one another, and (3) the DARPA-standard Sennheiser HMD-414 close-talking microphone. The subject wore the closetalking microphone and sat at a 1-meter distance from the other microphones. The signals from the electret microphones were passed through a filter with a response of –6 dB/octave between 125 Hz and 2 kHz, and a gain of 24 dB, to compensate for the frequency response of these microphones.

The training database for these experiments was from the original census data, obtained with a PZM6FS microphone with very different acoustical ambience. In order to compensate partially for differences between the training and environments, we normalized each cepstral coefficient (except for the zeroth) on an utterance-by-utterance basis by subtracting the mean of the values of that coefficient across all frames of the utterance.

Figure 4 shows the word error rates using cross-correlation processing with 1, 2, and 4 channels. It is seen that as more microphones are used, the word error rate decreases.We believe that the performance obtained using the PZM6FS microphones is better than that obtained with the electret microphones because PZM6FS microphones were used in the training database, indicating the need for a more effective type of environmental compensation than the simple mean normalization used in this pilot study.

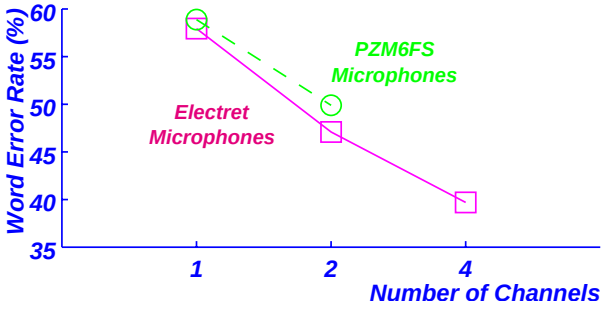


Figure 4: Comparison of word error rates for 1, 2 and 4 array elements using the electret microphones from the linear array and the PZM6FS omnidirectional desktop microphone.

In Table I we compare the results of 2-channel implementations of the cross-correlation algorithm (CC) with conventional LPC processing using the electret microphones from the linear array (ELECTRET) and the PZM6FS, along with single-channel processing of the Sennheiser closetalking microphone (CLSTLK). The feature set for recognition of the LPC-based system consists of 12 cepstral coefficients plus an additional power coefficient. These results indicate that for this more difficult database the cross-correlation processing is not yet producing a word error rate that is as good as the error rate obtained with conventional LPC processing, contrary to the results shown summarized in Figure 2. It is also surprising that the LPC-based performance using the electret microphone is better than that using the PZM6FS, as the PZM6FS was used to train the system. We believe that the performance of the 2-channel and 4-channel multi-microphone algorithms would be greatly improved by training on clean speech, better dynamic adaptation to new acoustical environments, better feature selection, and dynamic gain normalization.

In summary, the new multi-channel cross-correlation-based processing algorithm was found to preserve vowel spectra in the presence of additive noise and to provide greater recognition accuracy for the SPHINX-I speech recognition system than comparable processing of single-channel signals. Further increases in recognition accuracy should be obtained with the implementation of a small number of further design refinements.

.

MIC	LPC	CC
PZM6FS	57.4%	59.2%
CLSTLK	57.0%	59.3%
ELECTRET	47.5%	66.9%

Table I. Comparison of word error rates using a 2-channel implementation of the cross-correlation processing algorithm and conventional LPC-based processing, with three different microphones. The system was trained using the PZM6FS microphone.

ACKNOWLEDGMENTS

This research was sponsored by the Defense Advanced Research Projects Agency and monitored by the Space and Naval Warfare Systems Command under Contract N00039-91-C-0158, ARPA Order No. 7239.The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. We thank Robert Brennan and his colleagues at Applied Speech Technologies for consultations on their multichannel sampling hardware and software. We also thank the CMU speech group in general and Yoshiaki Ohshima in particular for many helpful conversations, ideas, and software.

REFERENCES

[1] Acero, A. and Stern, R. M., "Environmental Robustness in Automatic Speech Recognition", *ICASSP-90*, April 1990, pp. 849-852.

[2] Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G.W., "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms", *JASA*, Vol. 78, Nov. 1985, pp. 1508-1518.

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

[4] Peterson, P. M., "Adaptive Array Processing for Multiple Microphone Hearing Aids". RLE TR No. 541, Res. Lab. of Electronics, MIT, Cambridge, MA.

[5] Van Compernolle, D., "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings", *ICASSP-90*, April 1990, pp. 833-836.

[6] Lyon, R. F., "A Computational Model of Binaural Localization and Separation", *ICASSP-83*, 1983, pp. 1148-1151.

[7] Jeffress, L. A., "A Place Theory of Sound Localization", *J. Comp. Physiol. Psychol.*, Vol. 41, 1948, pp. 35-39.

[8] Stern, R. M., Jr., and Colburn, H. S., "Theory of Binaural Interaction Based on Auditory-Nerve Data. IV. A Model for Subjective Lateral Position", *J. Acoust. Soc. Amer.*, Vol. 64, 1978, pp. 127-140.

[9] Seneff, S., "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing", *Journal of Phonetics*, Vol. 16, No. 1, January 1988, pp. 55-76.

[10] Lee, K.F., *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.