ROBUST SPEECH RECOGNITION BASED ON HUMAN BINAURAL PERCEPTION

Richard M. Stern and Thomas M. Sullivan

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, based on human binaural hearing, 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, and compared to processing using simple delay-and-sum beamforming.

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.

In recent years there has been increased interest in the application of knowledge about signal processing in the human auditory system to improve the performance of automatic speech recognition systems (*e.g.* [2, 3, 4]). With some exceptions (*e.g.* [5, 6]), these algorithms have been primarily concerned with signal processing in the auditory periphery, typically at the level of individual fibers of the auditory nerve. While the human binaural system is primarily known for its ability to identify the locations of sound sources, it can also significantly improve the intelligibility of sound, particularly in reverberant environments [7]. In this paper we describe an algorithm that combines the outputs of multiple microphones to improve speech recognition accuracy. The form of this algorithm is motivated by knowledge of the more central processing that takes place in the human binaural system.

Since our algorithm processes the outputs of multiple microphones, it should be evaluated in comparison with other microphone-array approaches. Several types of array processing strategies have been applied to speech recognition systems. The simplest such system is the delay-and-sum beamformer (*e.g.* [8]). 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. This approach works reasonably well, but a relatively large number of microphones is needed for large processing gains. A second approach is to use an adaptive algorithm based on minimizing mean square energy, such as the Frost or the Griffiths-Jim algorithm [9]. 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.* [10]). (This problem can be avoided by only adapting during non-speech segments [11].)

The algorithm described in this paper is based on a third type of processing, 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.* [12]). 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.* [14, 16]).

In Sec. 2 we briefly review some aspects of human binaural processing, and we describe the new cross-correlation-based algorithm in Sec. 3. In Sec. 4 we describe typical results from pilot evaluations of the cross-correlation-based algorithm that demonstrate the algorithm's ability to preserve spectral contours. Finally, we describe in Sec. 5 the results of a small number of experiments that compare the speech recognition accuracy obtained with the new cross-correlation-based algorithm on conventional delay-and-sum beamforming.

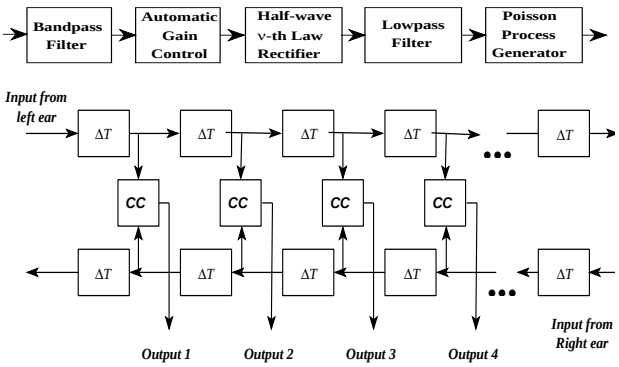


Figure 1. Upper panel: Block diagram of the transduction process in the auditory periphery. The output represents the response of a single fiber of the auditory nerve. Lower panel: Schematic representation of the Jeffress place mechanism. The blocks labelled ΔT indicate fixed timed delays in the signals.

2. CROSS-CORRELATION AND HUMAN BINAURAL PROCESSING

As a crude approximation, the peripheral auditory system can be characterized as a bank of bandpass filters, followed by some nonlinear post-processing. To the extent that such an offhand characterization is valid, we may further suggest that binaural interaction can be characterized as the cross-correlation from ear to ear of the outputs of peripheral channels with matching center frequencies [13].

Figure 1 is a schematic diagram of a popular mechanism that can accomplish the interaural cross-correlation operation in a physiologically-plausible fashion. This approach was originally proposed by Jeffress [14] and later quantified by Colburn [15] and others. The upper panel of Fig. 1 describes a functional model of auditory-nerve activity. This auditory-nerve model consists of (1) a bandpass filter to represent the frequency analysis performed by the auditory periphery, (2) a rectifier that represents nonlinearities in the transduction process, (3) a lowpass filter that represents the loss of synchrony of the auditory-nerve response to stimulus fine structure above about 1500 Hz, and (4) a mechanism that generates sample functions of a non-homogeneous Poisson process with an instantaneous rate that is proportional to the output of the rectifier.

The lower panel describes a network that performs temporal comparisons of the Poisson pulses arriving from peripheral auditory nerve fibers of the same characteristic frequency (CF), one from each ear, with successive delays of ΔT introduced along the path, as shown. The blocks labelled *CC* record coincidences of neural activity from the two ears (after the net delay incurred by the signals from the peripheral channels by the ΔT blocks). The response of a number of such units, plotted as a function of the net internal interaural delay can be thought of as an approximation to the interaural cross-correlation function of the sound impinging on the ear after the bandpass filtering, rectification, and lowpass filtering is performed by the auditory periphery. Figure 2 displays the relative

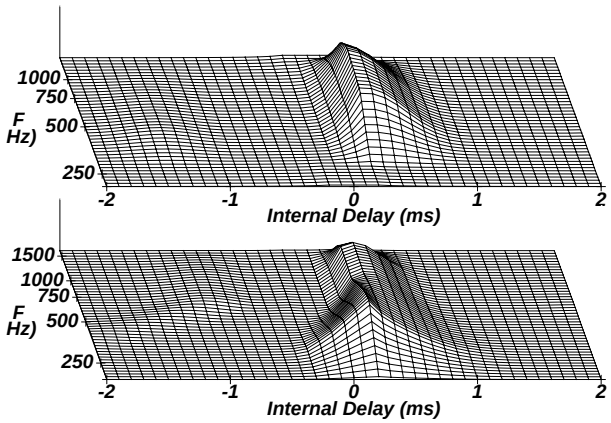


Figure 2. The response of an ensemble of binaural fiber pairs to a 500-Hz pure tone (upper panel) and to bandpass noise centered at 500 Hz (lower panel), each presented with a 0.5-ms ITD.

amount of activity produced by an ensemble of coincidence-counting units in response to two simple stimuli: a 500-Hz pure tone, and bandpass noise centered at 500 Hz, each presented with a 0.5-ms interaural time delay (ITD). The expected total number of coincidences is plotted as a function of internal delay (along the horizontal axis) and characteristic frequency (which is represented by the oblique axis). The diminished response for net internal delays greater than 1 ms in magnitude reflects the fact that only a small number of coincidence-counting units are believed to exist with those delays. In traditional binaural models, the location of the ridge along the internal-delay axis is used to estimate the lateral position or azimuth of a sound source. In this work we consider the spectral profile along the ridge (for more complex speech stimuli), and we specifically seek to determine the extent to which the cross-correlation processing of the binaural system serves to preserve the spectral contour along that ridge in difficult environments.

3. CROSS-CORRELATION-BASED MULTI-MICROPHONE PROCESSING

The goal of our multi-microphone processing is to provide a simplified computational realization of elements of the auditory system and of binaural analysis, but with potentially more than two sensors. In other words, we speculate what auditory processing might be like if we had 4, 8, or more ears. Figure 3 is a simplified block diagram of our 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. (This is the same processing performed by the conventional delay-and-sum beamformer.) 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. (The correlator outputs correspond to outputs of the coincidence counters at the internal delays of the "ridges" in Fig. 2.) Currently we use the 40-channel filterbank proposed by Seneff [2], which was designed to approximate the frequency selectivity of the auditory system. The shape of the

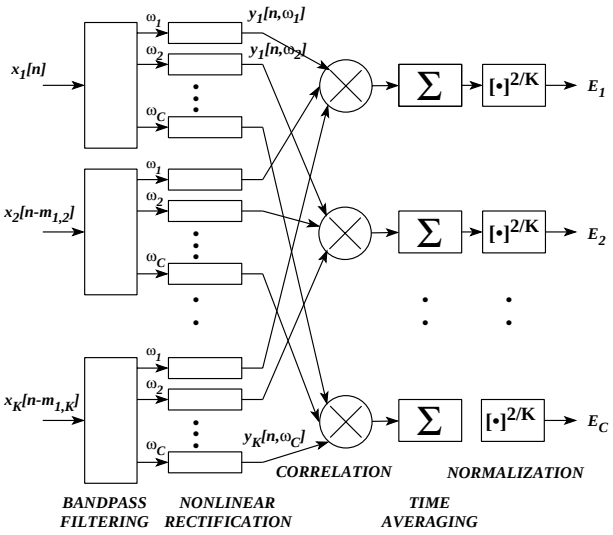


Figure 3. Block diagram of multi-microphone cross-correlation-based processing system.

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, these operations produce

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

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 12 cepstral coefficients using the cosine transform. The 12 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 [17].

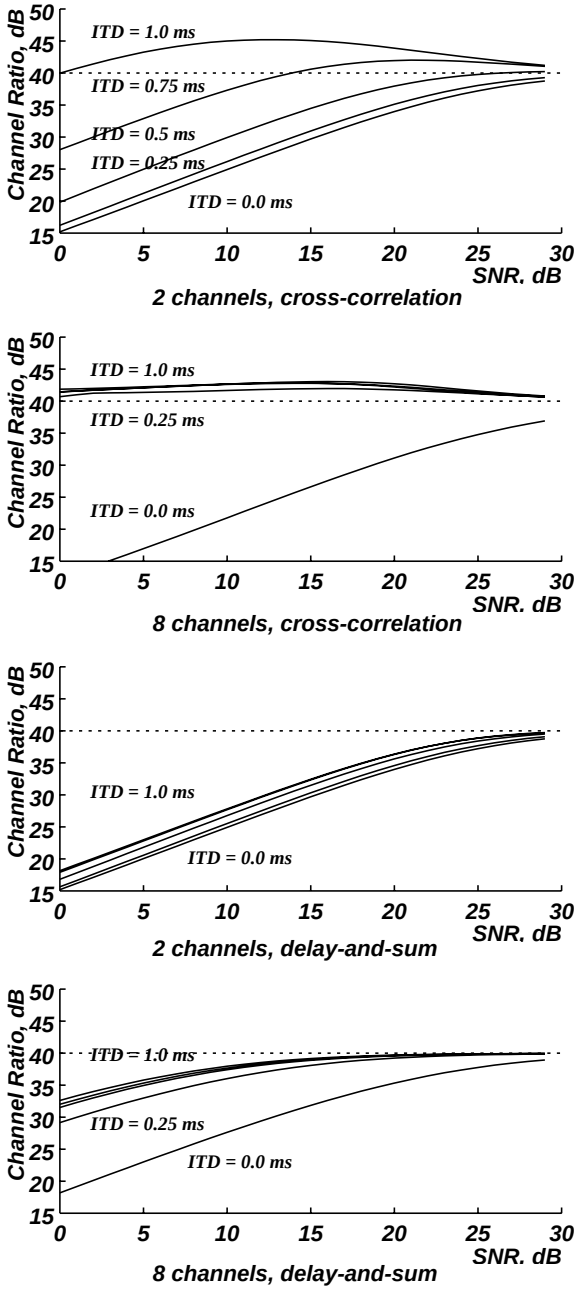


Figure 4. Comparisons of output energies of a 2-channel cross-correlation processor with delay-and-sum beamforming, using artificial additive noise. The actual power ratio of the two tones (without the noise) is 40 dB.

4. CROSS-CORRELATION PROCESSING AND ROBUST SPECTRAL PROFILES

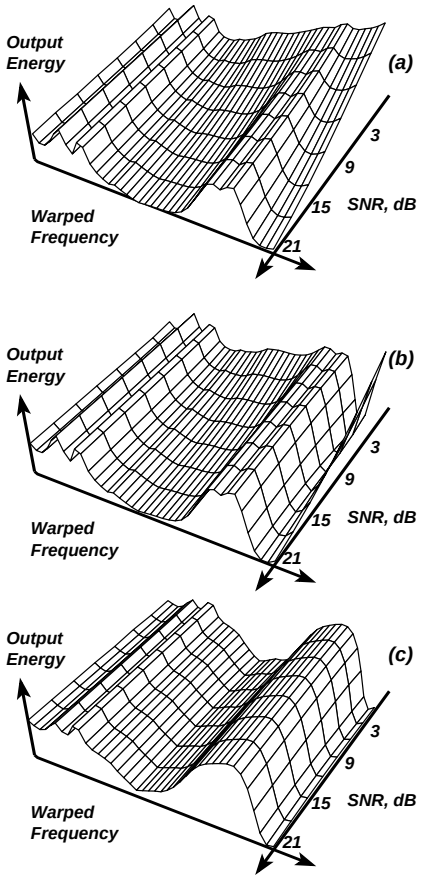
Comparisons using pairs of tones. We first evaluated the cross-correlation algorithm by implementing a series of pilot experiments with artificial stimuli. In the first experiment we examined the spectral profile developed by two sine tones, one at 1 kHz and

one at 500 Hz, with an amplitude ratio of 40 dB. The two tones were summed and corrupted by additive white Gaussian noise. The summed tones were presented identically to each "sensor" of the system (thus representing an "on-axis" signal), but the noise was added with a time delay from sensor to sensor that simulates the delay that is produced when the noise arrives at an oblique angle to a linear microphone array. Each sensor output was then passed through a pair of bandpass filters, one centered at 500 Hz and one at 1 kHz. The signals at the outputs of the bandpass filters were half-wave rectified, and the outputs from filters at corresponding frequency bands from each sensor were cross-correlated to extract an energy value for that frequency band. The ratio of these outputs was calculated and plotted for peak-signal-to-additive-noise ratios (SNR) ranging from 0 dB to 30 dB.

The results of this experiment are depicted in the four panels of Fig. 4, which display the power ratio of the outputs of the 500-Hz and 1000-Hz processing bands, as a function of SNR. In all cases, the ideal result would be the input power ratio of 40 dB, which is indicated by the horizontal dotted lines. Data were obtained for five values of sensor-to-sensor time delay (denoted "ITD"): 0.0, 0.25, 0.5, 0.75, and 1.0 ms. We compare results obtained using the cross-correlation array post processing as described above with processing in which the channels are summed prior to bandpass filtering. This case is representative of delay-and-sum beamforming, where the on-axis sine tone signal is reinforced relative to the off-axis uncorrelated noise signal. It can be seen in Fig. 4 that 8 sensors provides a better approximation than 2 sensors to the original 40-dB ratio of energies in the two frequency channels. For a given number of sensors, the cross-correlation algorithm performs better than delay-and-sum beamforming. Finally, with the desired signals presented simultaneously to the sensors, performance improves (unsurprisingly) as the sensor-to-sensor ITD of the noise is increased.

Comparisons using a synthetic vowel sound. We subsequently 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 (again, to simulate an off-axis corrupting signal impinging on a linear microphone 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 0. 125 ms to successive channels.

Figure 5 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



igure 5. Estimates of spectra for the vowel segment /a/ for varius 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.

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.

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

Encouraged by the appearance of these spectral profiles with simulated input, we evaluated 1-, 2-, 4-, and 8-channel implementations of the algorithm in the context of an actual speech recognition system. The CMU SPHINX-I speech recognizer [17] 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 2. All speech was sampled at 16 kHz. The frame size for analysis was 20 ms (320 samples) and frames were analyzed every 10 ms.

5.1. 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 ARPA 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 4 summarizes the results of these comparisons. Using the half-wave power-law rectifier with the positive signal raised to the 2^{nd} power (the "half-square" rectifier) provided the lowest word error rate of the various half-wave power-law rectifiers. The 2-channel cross-correlation algorithm provides a slightly better error rate than conventional LPC signal processing, and 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.

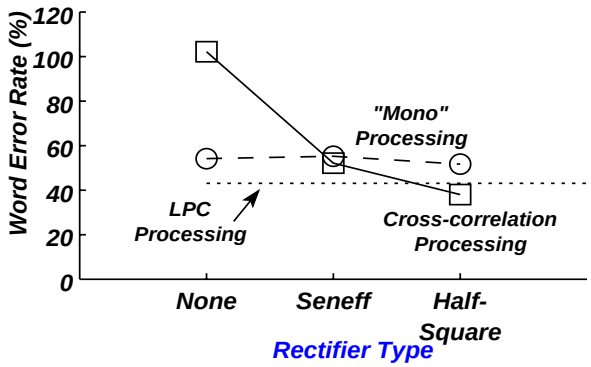


Figure 6. Comparison of word error rates achieved with 2-microphone processing using various half-wave rectifiers, and three types of signal processing.

5.2. Number of Processing Channels

We describe in this section results obtained using a new set of multiple-channel speech data. This testing database consisted of utterances from the CMU alphanumeric census task [1], and it 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 a single male speaker were collected using an 8-element linear array of inexpensive noise-cancelling pressure gradient electret condenser microphones, spaced 7 cm from one another. For comparison purposes, each utterance was also simultaneously recorded by a pair of omnidirectional desktop Crown PZM6FS microphones, also spaced 7 cm from one another, and the ARPA-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. By selecting a single element, the middle two elements, or the middle four elements from the 8-element array, arrays of 1, 2, 4, and 8 elements could easily be obtained.

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 7 shows the word error rates obtained using cross-correlation processing with 1, 2, 4, and 8 channels (microphones). The performance of three different algorithms is compared: (1) the original algorithm with auditory processing and the cross-correlation analysis (as in Fig. 3), (2) auditory processing used in conjunction with the initial delay-and-sum beamforming only, and (3) conventional LPC analysis in conjunction with simple delay-and-

sum beamforming. It is seen in each case that as more microphones are used, the word error rate decreases. The cross-correlation processing provides lower error rates for the 2- and 4-microphone cases, but all 3 methods give roughly the same performance for the 8-microphone case.

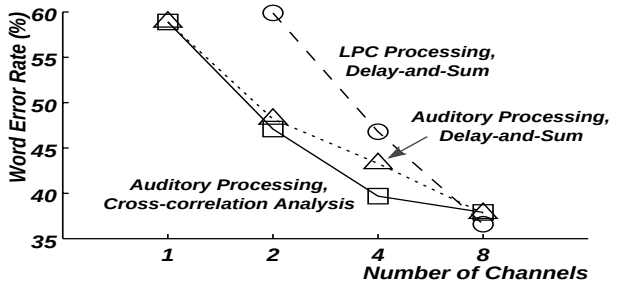


Figure 7. Comparison of word error rates for 1, 2, 4, and 8-channel array processors using the electret microphones of the Flanagan array. The system was trained on speech using the PZM6FS microphone. Three types of processing are compared: auditory-based pre-processing using delay-and-sum beamforming and the cross-correlation-based enhancement (boxes), auditory-based pre-processing using delay-and-sum beamforming alone (triangles), and LPC processing using delay-and-sum beamforming alone.

6. 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 compared to comparable processing of single-channel signals, and compared to comparable processing using delay-and-sum beamforming in the cases examined. We expect to observe further increases in recognition accuracy as further design refinements are introduced to the algorithm.

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, by NSF Grant IBN 90-22080, and by the Motorola Corporation, which has supported Thomas Sullivan's graduate research. 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 multi-channel sampling hardware and software. We also thank the CMU speech group in general and Yoshiaki Ohshima in particular for many helpful conversations, good ideas, and software packages.

REFERENCES

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

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

3. Lyon, R. F., "A Computational Model of Filtering, Detection, and Compression in the Cochlea", ICASSP-82, pp. 1282-1285, 1982.

4. Ghitza, O.,"Auditory Nerve Representation as a Front-End for Speech Recognition in a Noisy Environment", *Comp. Speech and Lang*, **1,** pp. 109-130**,** 1986**.**

5. Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M., "Complex Sounds and Auditory Images", *Auditory Physiology and Perception*, Cazals, Y., Horner, K., and Demany, L.", Eds., pp. 429-446, Pergamon Press, 1991.

6. Duda, R.O.; Lyon, R.F.; Slaney, M., Correlograms and the Separation of Sounds, *Proc. Twenty-Fourth Asilomar Conference on Signals, Systems and Computers*, **1,** pp. 457-46, Maple Press, 1990.

7. Blauert, J., "Binaural Localization: Multiple Images and Applications in Room- and Electroacoustics", *Localization of Sound: Theory and Applications*, R. W. Gatehouse, Ed., pp. 65-84, Amphora Press, Groton CT, 1982.

8. 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.

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

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

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

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

13. Stern, R. M., and Trahiotis, C., "Models of Binaural Interaction", *Handbook of Perception and Cognition, Volume 6: Hearing*, B. C. J. Moore, Ed., Academic Press, 1995.

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

15. Colburn, H. S., "Theory of Binaural Interaction Based on Auditory-Nerve Data. I. General Strategy and Preliminary

Results on Interaural Discrimination", *J. Acoust. Soc. Amer.*, **54**, pp. 1458-1470", 1973.

16. 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.

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