"Polyaural" Array Processing for Automatic Speech Recognition in Degraded Environments

Richard M. Stern, Evandro Gouvêa, and Govindarajan Thattai

Department of Electrical and Computer Engineering and Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213 USA

rms@cs.cmu.edu, egouvea@cs.cmu.edu,govind@cs.cmu.edu

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 halfwave rectification operations, and then cross-correlating the outputs from each channel within each frequency band. These operations provide rejection of off-axis interfering signals. These operations are repeated (in a non-physiological fashion) for the negative of the signal, and an estimate of the desired signal is obtained by combining the positive and negative outputs. We demonstrate that the use of this approach provides substantially better recognition accuracy than delay-and-sum beamforming using the same sensors for target signals in the presence of additive broadband and speech maskers. Improvements in reverberant environments are tangible but more modest.

Index Terms: robust speech recognition, binaural hearing, auditory processing, speech enhancement

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. 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.

In recent years, the use of arrays of microphones has become increasingly popular as a means to improve automatic speech recognition accuracy in situations where a signal and competing noise sources are spatially separated. Several different types of array processing strategies have been applied to speech recognition systems. The simplest such system is the delay-and-sum beamformer as used by the work of Flanagan and his colleagues (*e.g.* [1]). 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 [2]. 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 generally 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. The LIMABEAM algorithm developed by Seltzer *et al.* [3] represents an interesting new approach to optimal array processing for speech recognition in which the objective of the adaptation is to minimize distortion of the features used in speech recognition, rather than waveform distortion. A variant of this algorithm that uses subband processing [4] has demonstrated some improvement in speech recognition accuracy in reverberant environments.

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 in recent years an increasing number of researchers have developed signal processing strategies for speech recognition systems that are based on human binaural processing (*e.g.* [5, 6, 7]). These approaches, as well as others reviewed in [8] typically use short-time Fourier transformation to decompose incoming speech into components that are localized in time and frequency and subsequent binaural analysis to determine which time-frequency components are most likely to belong to the target speaker. The enhanced speech is then obtained by performing short-time Fourier synthesis only on the components of the input that are likely to be dominated by the desired signal. The algorithm described in this paper, on the other hand, relies on nonlinear processing motivated by the auditory system to obtain an enhanced representation of the desired signal and reconstructs the enhanced signal using *all* time-frequency components of the input.

We describe our new cross-correlation-based algorithm in the following section, and we describe the impact of the algorithm on automatic speech recognition accuracy in Sec. 3.

2. Polyaural processing

The human binaural system is well known for its ability to localize and separate sound sources according to their direction of arrival, as well as for its ability to improve the intelligibility of speech signals in reverberant environments [9]. Because of these extraordinary capabilities, many useful characterizations have been developed to describe how the binaural system oper-

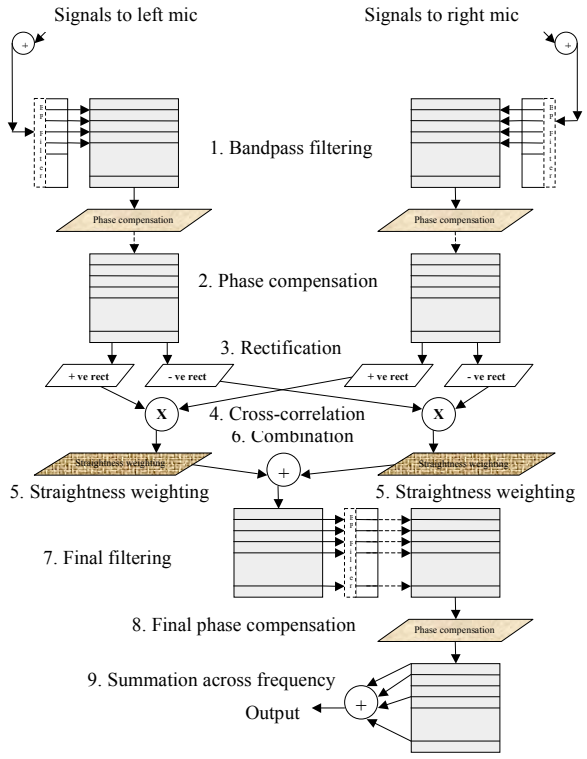

Figure 1: *Schematic diagram of polyaural processing. Although only two microphones are depicted, the processing is extensible to an arbitrary number of microphones.*

ates (*e.g.* [10, 11]). To begin, signals entering the ears are first processed by the peripheral auditory system, which is typically modeled by a bank of bandpass filters followed by nonlinear processing of the outputs of the filters that includes halfwave rectification. Binaural processing is typically modeled as an interaural cross-correlation of the outputs of peripheral auditory fibers that are matched in terms of their best frequency of response (or in other words, an interaural cross-correlation of matching channels of the bandpass filterbank). Some researchers also argue that a second level of cross-correlation is also performed across frequency, which serves to emphasize those components of the binaural response that are consistent over frequency or "straight" [12]. Reviews such as [10] discuss in detail how these representations are useful for various attributes of classical binaural processing that are studied in the psychophysical literature.

In developing the processing algorithm for the present work we were interested in obtaining a representation that would include the normal binaural processing, but do so in a fashion that is extensible to more than two "ears", and that would enable us to recover a continuous waveform. We refer to this approach as "polyaural", or extended binaural processing. Polyaural processing is accomplished by the process that is summarized in Fig. 1. Specifically, the incoming signals undergo the following stages of processing:

1. The signals to the left and right mics (which generally contain multiple components with different delays from mic to mic) are passed through a bank of bandpass filters. We used conventional Gammatone filters for this purpose as implemented in Slaney's auditory toolbox [13].
2. The phase differences of the filters are compensated for by time-reversing the original filter outputs, passing them through the filter bank again, and time-reversing the final outputs. (This is equivalent to convolving the original filter outputs with the same filter in time-reversed form, effectively providing zero-phase filtering in each channel.)
3. The phase-compensated outputs are passed through a half-wave rectifier that has zero response for negative input and that raises its input to the ν^{th} power when the input is positive. ν is typically a small integer. The phase-compensated filter outputs are also negated and passed through similar half-wave rectifiers in parallel channels.
4. The positive and negative rectifier outputs are cross-correlated separately. This is accomplished by imposing a delay to compensate for the relative propagation delay (if any) of the desired signal to the mics and then multiplying across channels.
5. The cross-correlated outputs are *optionally* correlated a second time across a limited range of frequencies to provide the straightness weighting. Regardless of whether or not straightness weighting is employed, the resulting outputs are passed through a nonlinearity of power $1/NM\nu$ where N is the number of microphones, M is the number of straightness weighting channels, and ν is the order of the rectification imposed in Step 3. The intended effect is that of restoring the desired signal to its original value.
6. The positive and negative components are added together.
7. The output signals are passed through a second bank of bandpass filters.
8. The final filters are phase compensated as before.
9. The resulting signals are combined across frequency.

Steps 1 and 3 are intended to model the auditory periphery (in a rather crude fashion), with Step 2 added to maintain phase coherence over frequency for the optional straightness weighting in Step 5. Step 4 represents the interaural correlation that is a major feature of many models of binaural interaction. The halfwave processing of the negative portions of the signals described in Steps 3 through 6 and depicted in the right side of Fig. 1 for these steps is grossly non-physiological, of course. This processing has been included to permit the reconstruction of a complete waveform with both positive and negative values. The second bandpass filtering and phase compensation in Steps 8 and 9 are included to attenuate the many spurious distortion components that are introduced by the discontinuities in the derivatives of off-axis signal components produced by the half-wave rectification and cross-correlation.

The phase compensation of the peripheral filters in Step 2, while not needed in theory if straightness weighting is not used, has been found to be quite useful in practice, most likely because the filters have overlapping frequency responses, and phase compensation reduces the likelihood of uncontrolled destructive or constructive interference among adjacent frequency channels.

This series of operations, if executed properly, has the effect of leaving the desired signal intact if it is presented in isolation. Signals arriving from the side, on the other hand, tend to be attenuated because the disparity in arrival times causes the time

periods during which both signals are positive (or negative) to be limited, which reduces the output of the cross-correlation.

As an example of the efficacy of this processing, we provide in the Proceedings examples of speech enhanced by polyaural processing using a generalization to the two-mic implementation described above in the form of an array of 11 logarithmically-spaced sensors in the manner of Flanagan *et al.* [1], as described in Sec. 3 below. The signals are digitally combined in a fashion that corresponds to a dominant source arriving from a direction along the perpendicular bisector to the array, and a second source arriving at an angle of approximately 45 degrees to one side. The examples compare delay-and-sum processing, and polyaural processing with and without straightness weighting. These examples are also available online [14].

3. Experimental Results

We describe in this section the results of initial experiments intended to assess the extent to which the polyaural processing described in the previous section can improve speech recognition accuracy beyond the level of accuracy obtained using simple delay-and-sum beamforming. We first describe the procedures that were used in all experiments and then discuss our initial recognition results with additive noise and in simulated reverberant environments.

3.1. Experimental Procedures

Speech recognition was measured using the well known DARPA Resource Management (RM1) database using the Carnegie Mellon Sphinx-3 system, which is available in open source form at `http://cmusphinx.org`.

The system was implemented using 3-state continuous HMMs, 1000 senones, and 8 Gaussian mixtures for the output densities. The models were trained using a subset of the speaker independent portion of the RM1 database containing 1600 utterances, recorded in a clean environment using a close-talking microphone at a sampling rate of 16 kHz. Evaluation results were obtained using a subset of the speaker-independent RM1 test set containing 600 randomly-selected utterances with a total of 5681 words.

The environments used in the present study were digitally simulated, with the input device presumed to be an 11-element logarithmic array of the type proposed by Flanagan [1]. This array consisted of four nested 5-element arrays with inter-element spacings of 3 cm, 6 cm, 9 cm, and 12 cm, respectively, using shared microphones where possible. As discussed in [1], these sub-arrays each process input in different frequency bands, enabling the beamwidth in the look direction to remain more constant over a wide range of frequencies than would be possible with a simple linear array.The target speaker was assumed to be standing along the perpendicular bisector of the line defined by the array elements, and the interfering source (when there was one) was assumed to be located at an azimuth of approximately 45 degrees. Consequently, the sample delay between the center microphone and its closest neighbors is one sample.

3.2. Performance in the presence of additive disturbances

Using the simulated physical topology described above, we measured the word error rate (WER) obtained when an interfering source was white noise or when it was a second speaker. The specific interfering speaker was selected at random, so the target and interfering speakers were of the same gender for some but not all of the trials. The white noise samples used were obtained from the Noisex-92 database.

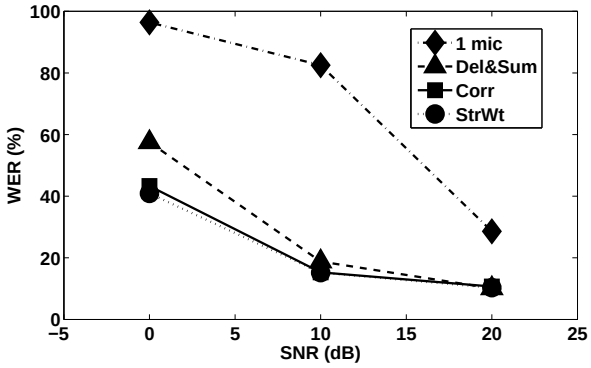

Figure 2: *Speech recognition accuracy in the presence of interfering noise as a function of SNR. The target speech is at an azimuth of 0 degrees relative to the normal to the plane of the array, and the interfering source arrives at approximately 45 degrees. Percentage WER is depicted for a single microphone (diamonds), delay-and-sum beamforming (triangles), direct correlation processing (squares), and correlation processing with "straightness" weighting (circles). See text for details.*

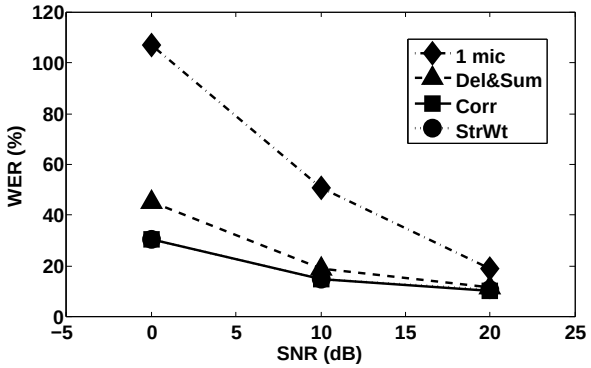

Figure 3: *Same as Figure 2, except that WER is measured in the presence of speech maskers.*

The target speech and the interfering signals were combined at SNRs of 0, 10, and 20 dB, as measured directly from the energies of the target and interference. In our experiments, the interfering source had little effect on recognition accuracy for SNRs above 20 dB with the array processing.

Figure 2 depicts the WER obtained using four types of processing: a single omnidirectional microphone (diamonds), a simple Flanagan delay-and-sum array (triangles), the physiologically-motivated cross-correlation processing without weighting for "straightness" (squares), and the cross-correlation processing with the additional straightness weighting (circles). As has been reported previously, array processing provides a dramatic improvement compared to processing with a single microphone, even in the simple delay-and-sum configuration. Nevertheless, the cross-correlation based processing provides a relative improvement in WER of about 18.6 percent at 10 dB SNR and 24.7 percent at 0 dB. Stated another way, the use of the cross-correlation processing provides an effective improvement in SNR of roughly 2-4 dB at SNRs of 0 to 5 dB. The straightness weighting appears to have little effect on the results, at least for these data.

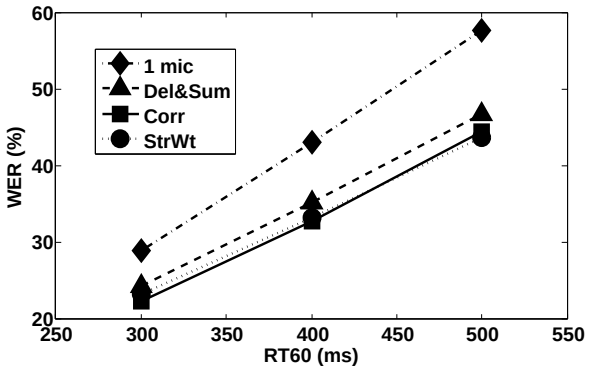

Figure 4: *Similar to the previous two figures, except that WER is measured in simulated reverberant environments with the source and array separated by 2 m. See text for details.*

Figure 3 depicts similar results, except that speech maskers from the RM1 database are used rather than the noise maskers in Fig. 2. As in the case with the noise maskers, the microphone array processing is quite effective in separating the sources and the correlation-based processing provides even greater decreases in relative WER, approximately 22.0 and 32.3 percent at 10 dB and 0 dB, respectively.

3.3. Performance in reverberant environments

We also examined the ability of the correlation-based processing to provide substantial gains in recognition accuracy in very difficult reverberant environments. We simulated the effects of room reverberation using the image method [15] using the publicly-available package `rir` which can be downloaded from `http://2pi.us/rir.html`.

We simulated a room with dimensions 5m x 4m x 3m (W x L x H), with the microphone array located exactly in the middle of the room, perpendicular to the width. The speaker is located perpendicular to the array, at distances of 1 and 2 meters from the array, and the uniform reflectance of the surfaces of this "shoebox" model of a room was manipulated to provide reverberation times of 300 to 500 ms.

Figure 4 displays representative results from these simulations, with 2 meters separating the source and microphones. Polyaural provided more modest but decreases in relative error rate of 7.1 percent and 6.4 percent at reverberation times of 400 and 500 ms, respectively. Trends at the 1-m distance were similar but differences between conditions were smaller in magnitude.

4. Discussion and Conclusions

In this paper we have introduced the "polyaural processing" method of improving speech recognition accuracy in the presence of spatially-separated interfering sources and in reverberant environments. In this procedure, parallel frequency channels are processed in a nonlinear fashion to enhance the desired signal and suppress interfering components from other directions. The proposed method provides very substantial improvements in recognition accuracy compared to baseline delay-and-sum processing in the presence of interfering speech and broadband sources. Improvements in accuracy in reverberant environments are also tangible but more modest.

5. Acknowledgements

This research was supported by the National Science Foundation (Grant IIS-0420866).

6. References

[1] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoustic. Soc. Amer.*, vol. 78, pp. 1508–1518, 1985.

[2] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Prentice-Hall, 1985.

[3] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, September 2004.

[4] M. L. Seltzer and R. M. Stern, "Subband likelihood-maximizing beamforming for speech recognition in reverberant environments," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2109–2121, November 2006.

[5] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, no. 4, pp. 361–378, 2004.

[6] N. Roman and D. L. Wang, "Binaural tracking of multiple moving sources," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. V, 2003, pp. 149–152.

[7] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[8] R. M. Stern, D. Wang, and G. Brown, "Binaural sound localization," in *Computational Auditory Scene Analysis*, G. Brown and D. Wang, Eds. Wiley and IEEE Press, 2006.

[9] J. Blauert, *Spatial Hearing*. Cambridge, MA: MIT Press, 1997, revised edition.

[10] R. M. Stern and C. Trahiotis, "Models of binaural interaction," in *Hearing*, ser. Handbook of Perception and Cognition, B. C. J. Moore, Ed. Academic (New York), 1995, ch. 10, pp. 347–386.

[11] H. S. Colburn and A. Kulkarni, "Models of sound localization," in *Sound Source Localization*, ser. Springer Handbook of Auditory Research, R. Fay and T. Popper, Eds. Springer-Verlag, 2005, ch. 8, pp. 272–316.

[12] R. M. Stern and C. Trahiotis, "The role of consistency of interuaral timing over frequency in binaural lateralization," in *Auditory physiology and perception*, Y. Cazals, K. Horner, and L. Demany, Eds. Pergamon Press, Oxford, 1992, pp. 547–554.

[13] M. Slaney, *Auditory Toolbox (V.2)*, 1998. [Online]. Available: http://www.slaney.org/malcolm/pubs.html

[14] R. M. Stern, E. Gouvêa, and G. Thattai, *Examples of polyaural processing*. [Online]. Available: http://www.cs.cmu.edu/∼robust/Papers/Polyaural0807.tar.gz

[15] J. B. Allen and D. A. Berkley, "Image method for efficieintly simulating small room acoustics," *J. Acoustic. Soc. Amer.*, vol. 65, pp. 943–950, 1979.