BINAURAL AND MULTIPLE-MICROPHONE SIGNAL PROCESSING MOTIVATED BY AUDITORY PERCEPTION

Richard M. Stern, Evandro Gouvêa, Chanwoo Kim, Kshitiz Kumar, and Hyung-Min Park

Department of Electrical and Computer Engineering and Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213 USA
rms@cs.cmu.edu

ABSTRACT

It is well known that binaural processing is very useful for separating incoming sound sources as well as for improving the intelligibility of speech in reverberant environments. This paper describes and compares a number of ways in which the classic model of interaural cross-correlation proposed by Jeffress, quantified by Colburn, and further elaborated by Blauert, Lindemann, and others, can be applied to improving the accuracy of automatic speech recognition systems operating in cluttered, noisy, and reverberant environments. Typical implementations begin with an abstraction of cross-correlation of the incoming signals after nonlinear monaural bandpass processing, but there are many alternative implementation choices that can be considered. These implementations differ in the ways in which an enhanced version of the desired signal is developed using binaural principles, in the extent to which specific processing mechanisms are used to impose suppression motivated by the precedence effect, and in the precise mechanism used to extract interaural time differences.

Index Terms – binaural hearing, robust speech recognition, reverberation, auditory models

1. INTRODUCTION

We listen to speech (as well as to other sounds) with two ears, and it is quite remarkable how well we can separate and selectively attend to individual sound sources in a cluttered acoustical environment. In fact, the familiar term "cocktail party processing" was coined in an early study of how the binaural system enables us to selectively attend to individual conversations when many are present, as in, of course, a cocktail party. This phenomenon illustrates the important contribution that binaural hearing makes to auditory scene analysis, by enabling us to localize and separate sound sources. In addition, the binaural system plays a major role in improving speech intelligibility in noisy and reverberant environments.

In this paper we discuss some of the ways in which the known characteristics of binaural processing have been exploited in recent years to separate and enhance speech signals, and to improve automatic speech recognition accuracy in difficult acoustical environments. Like so many aspects of sensory processing, the binaural system offers an existence proof of the possibility of extraordinary performance in sound localization and signal separation, but it does not yet provide a very complete picture of how this level of performance can be achieved with the tools available in contemporary signal processing.

Supported in part by NSF Grant IIS-0420866.

2. ASPECTS OF BINAURAL PERCEPTION

2.1. Physical cues

A number of factors affect the spatial aspects of how a sound is perceived. As Rayleigh noted [1], two physical cues dominate the perceived location of an incoming sound source. An *interaural time difference* (ITD) is produced because it takes longer for the sound to arrive at the ear that is farther from the source. The signal to the ear closer to the source is also more intense because of the "shadowing" effect of the head, producing an *interaural intensity difference* (IID). IIDs are most pronounced at frequencies above approximately 1.5 kHz because it is only at these frequencies that the head is large enough to reflect the incoming sound wave. ITDs exist at all frequencies, but periodic sounds can be decoded unambiguously only for frequencies for which the maximum physically-possible ITD is less than half the period of the waveform at that frequency, or at frequencies below 1.5 kHz for typically-sized human heads.

2.2. Physiological estimation of ITD and IID

Binaural interaction, of course, takes place after peripheral processing that includes the frequency selectivity of the cochlea and the transduction of mechanical motion in the cochlea to electrical impulses that are transmitted along the fibers of the auditory nerve. Many computational models have been developed to describe these phenomena (*e.g.* [2, 3, 4]), typically incorporating bandpass filtering (with the frequency of best response in each channel referred to as the *characteristic frequency* or CF), nonlinear rectification, and other phenomena such as saturation and lateral suppression. The neural response to frequency components up to about 1.5 kHz are synchronized to the cycle-by-cycle timing information of the incoming signals, which enables the estimation of ITDs.

A number of cells in the brainstem are likely to be useful in extracting the ITDs and IIDs used in auditory spatial perception, as reviewed by [5, 6], among others. One of the most significant physiological findings has been the observation of cells that appear to detect specific ITDs, independent of frequency, first reported by Rose *et al.* [7] in the the brainstem. This delay is referred to as a *characteristic delay* (CD). The results of must studies suggest that ITD-sensitive cells tend to exhibit CDs that lie in a broad range of ITDs, with the density of CDs decreasing as the absolute value of the ITD increases. In a recent series of studies McAlpine and his collaborators have argued that most ITD-sensitive units exhibit characteristic delays that occur in a narrow range that is close to

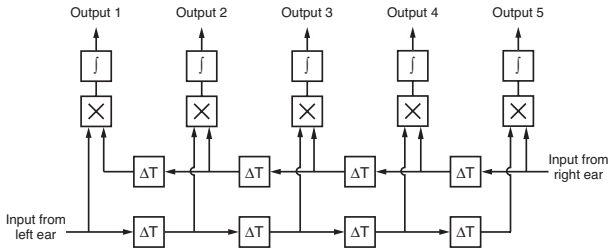

Fig. 1. Schematic representation of the Jeffress-Colburn model. Boxes containing crosses are correlators (multipliers) that record coincidences of neural activity from the two ears after the internal delays (ΔT). This structure is the basis for the models proposed by many others.

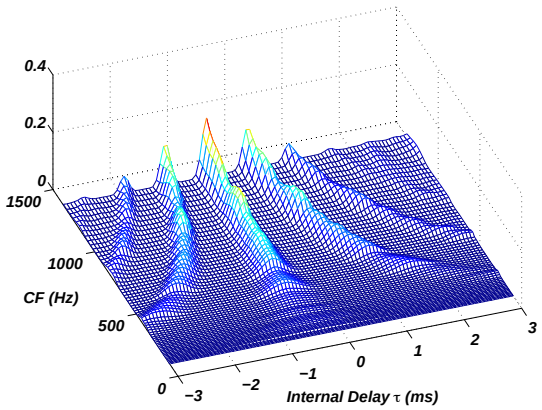

Fig. 2. Representation of bandpass noise by the Jeffress-Colburn model. The upper panel of the figure shows the rate of response of individual Jeffress-Colburn coincidence detectors to bandpass noise presented with an ITD of -0.5 ms. The relative rate of coincidences of each fiber pair are depicted as a joint function of CF and internal delay.

approximately one-eighth of the period of a cell's characteristic frequency (*e.g.*) [8]), at least for some animals.

2.3. Some binaural phenomena

The human binaural system is remarkable in its ability to localize single and multiple sound sources, to separate and segregate signals coming from multiple directions, and to understand speech in noisy and reverberant environments. There have been many studies of binaural perceptual phenomena, and useful comprehensive reviews may be found in [9] and [10], among other sources, and some results relevant to robust speech recognition are reviewed more recently in [11].

While the scope of this paper does not permit a comprehensive review of binaural phenomena, a small number of major results that are especially relevant to this discussion include: (1) the perceived laterality of sound sources depends on both the ITD and IID at the two ears, although the relative salience of these cues depends on frequency. (2) The auditory system is exquisitely sensitive to small changes and sound, and can discriminate ITDs on the order of 10 μs and IIDs on the order of 1 dB. Sensitivity to small differences in interaural correlation of broad-band noise sources is also quite acute, as a decrease in interaural correlation from 1.00 to 0.96 is readily discernible. (3) The vertical position of sounds, as well as front-to-back differentiation in location, is affected by changes in the frequency response of sounds that are imparted by the outer ear, and reinforced by head-motion cues. (4) The intelligibility of speech in noise is greater if the interaural differences of the target are different from those of the masker. Some of this improvement can be attributed to the simple fact that one of the ears has a greater effective SNR than the average, but binaural interaction also plays a significant role (*e.g.* [12, 13]).

It has long been noted that in a reverberant environment the auditory localization mechanisms pay greater attention to the first component that arrives (which presumably comes directly from the sound source) at the expense of the latter-arriving components (which presumably are reflected off the room and/or objects in it. This phenomenon is referred to as the *precedence effect* or the *law of the first wavefront*. Blauert and others have noted that the precedence effect is likely to play an important role in increasing speech intelligibility in reverberant environments (*e.g.* [14]).

2.4. Models of binaural interaction

Most modern computational models of binaural perception are based on Jeffress's description of a neural "place" mechanism that would enable the extraction of interaural timing information [15]. Jeffress postulated a mechanism that consisted of a number of central neural units that recorded coincidences in neural firings from two peripheral auditory-nerve fibers, one from each ear, with the same CF. He further postulated that the neural signal coming from one of the two fibers is delayed by a small amount that is fixed for a given fiber pair, as in the block diagram of Fig. 1. Because of the synchrony in the response of low-frequency fibers to low-frequency stimuli, a given binaural coincidence-counting unit at a particular frequency will produce maximal output when the external stimulus ITD at that frequency is exactly compensated for by the internal delay of the fiber pair. Hence, the external ITD of a simple stimulus could be inferred by determining the internal delay that has the greatest response over a range of frequencies.

Colburn [16] reformulated Jeffress's hypothesis quantitatively using a relatively simple model of the auditory-nerve response to sound, and a "binaural displayer" consisting of a matrix of coincidence-counting units of the type postulated by Jeffress. These units are specified by the CF of the auditory-nerve fibers that they receive input from as well as their intrinsic internal delay. If the duration of the coincidence window is sufficiently brief, it can be shown that at each CF the pattern of activity developed by these coincidence-counting units is approximately the cross-correlation function of the neural response to the signals to the ears at that frequency (after the peripheral auditory processing).

There have been a number of subsequent enhancements proposed for the basic Jeffress-Colburn model. For example, Stern and Colburn describe a mechanism that predicts subjective lateral position based on ITD and IID [17]. Lindemann [18], extending earlier work of Blauert [19], added a mechanism that inhibits outputs of the coincidence counters when there is activity produced by coincidence counters at adjacent internal delays, and introduced a

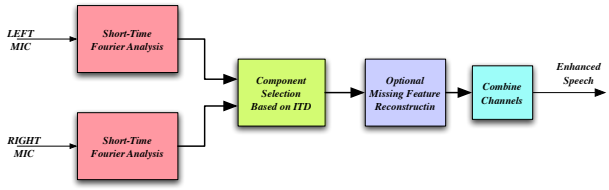

Fig. 3. *Schematic diagram of zero-crossing-based amplitude estimation (ZCAE) processing. See text for details.*

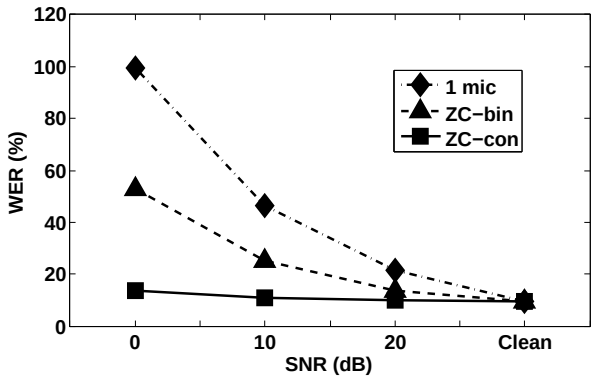

Fig. 4. *Speech recognition accuracy using the ZCAE algorithm in the presence of an interfering speech source as a function of SNR in the absence of reverberation. Percentage WER is depicted for a single microphone (diamonds), ZCAE using binary decision-making (triangles), and ZCAE using continuous estimate of target probability. See text for details about the input.*

monaural-processing mechanisms at the "edges" of the display of coincidence-counter outputs that become active when the intensity of the signal to one of the two ears is extremely small. The contralateral inhibition mechanism enables the Lindemann model to describe several interesting phenomena related to the precedence effect [20]. Gaik [21] extended the Lindemann mechanism further by adding a second weighting to the coincidence-counter outputs that reinforces naturally-occurring combinations of ITD and IID. Stern and Trahiotis [22] proposed a secondary network that recorded coincidences across frequency at each ITD, which reinforces components of the representation that are consistent across frequency.

Figure 2 depicts the relative number of coincidences of the original Jeffress-Colburn model that are developed in response to a bandpass-noise signal presented with an ITD of -0.5 ms, plotted as a joint function of CF and internal delay. It can be seen that the true delay of the signal is indicated by a vertical ridge in the plot at the internal delay of -0.5 ms. We also note that the ridges are somewhat broad along the internal-delay axis. A natural sharpening of these ridges occurs when either the contralateral inhibition proposed by Lindemann [18] or the second-level of coincidences over frequency proposed by Stern and Trahiotis [22] is added to the basic Jeffress-Colburn model. Several computational models accomplish this task by first estimating the putative ITD and subsequently developing a "skeletonized" cross-correlation function in which the peaks of the original model are replaced by relatively narrow Gaussian-shaped ridges (*e.g.* [23, 24]).

3. APPLICATION TO ROBUST SPEECH RECOGNITION

There has been a great deal of interest over the past two decades in the application of knowledge of binaural processing to improvements in the performance of automatic speech recognition systems. In this section we describe a small sample of such systems, with regrets that limitations of space preclude a more comprehensive listing of technologies and results. A more comprehensive summary of many of these techniques may be found in [25].

3.1. Early approaches

The first application of binaural modeling to automatic speech recognition was by Lyon [26], who combined an auditory model with a computational model of binaural processing based on the Jeffress model, segregating the desired signal according to ITD. The model was evaluated only subjectively; it was reported to show improvement in "dry" environments, and less improvement in the presence of reverberation.

Several systems based on various implementations of the systems developed by Blauert, Lindemmann and Gaik were developed in the 1990s including the "cocktail-party processor" described by Bodden [27] which includes the computational model of Lindemann with contralateral inhibition and the enhancements by Gaik which weight more heavily signal components with plausible combinations of ITD and IID. Bodden and Anderson [28] later described an effective improvement of 20 dB in SNR through the use of the Bodden processor and other enhancements for simulated speech arriving on axis in the presence of noise at a 30-degree angle. The stimuli in these experiments were generated digitally with no attempt to incorporate a model of reverberation.

3.2. Selective reconstruction based on binaural analysis

In the past decade a large number of systems have been developed and evaluated that use principles of computational auditory scene analysis (CASA) and missing-feature reconstruction. These systems typically analyze incoming speech signals from cluttered and potentially reverberant acoustical environments to identify those components of the input which are dominated by the target signal. A "mask" is developed that separates the desired input components from those that are believed to be dominated by noise, distortion, or interfering sources. The systems then "selectively reconstruct" the desired speech waveform based only on these "good" components from the input or they develop features to represent it based on the same subset of the input. Controlled evaluation of these systems in conditions that approximate reverberant environments has been greatly facilitated by the development and widespread availability of room impulse response simulations based on the image method [29] such as `RIR` [30] and `ROOMSIM` [31].

Many interesting systems were developed through a long series of collaborations between researchers at Ohio State University and the University of Sheffield. For example, the system of Roman *et al.* [32] localized targets in reverberant environments based on ITD and then determined which frequency components at that ITD are dominated by target components based on empirical observations of the ITD and IID. Palomäki et al. [24] elaborated on that approach by adding a mechanism proposed by Martin [33] to model the precedence effect before binaural processing. Martin's precedence mechanism emphasizes the transient segments of incoming signals, which are more likely to be in the direct field. The system also incorporated the "skeletonized" abstraction of the cross-

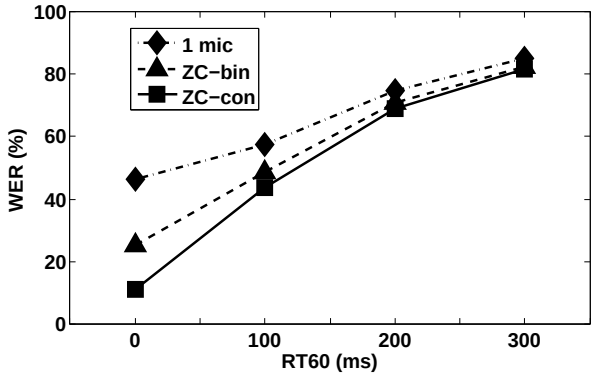

Fig. 5. *Similar to the previous figure, except that WER is measured in simulated reverberant environments at an SNR of 10 dB with the source and array separated by 2 m. See text for details.*

correlation function as described above and a much more sophisticated mask-estimation procedure that considered IID, temporal modulation, and energy in determining which components in the representation are likely to be dominated by the target. Roman *et al.* [34] obtained still better performance with a 2-microphone system that used adaptive filtering to cancel the target signal as part of the mask-estimation processing. Srinivasan *et al*. [35] developed several types of ratio masks (representing the relative dominance of the target as a continuous rather than a binary function) using a time-varying Wiener filter based on empirically-observed combinations of ITD and IID developed by natural stimuli.

Park developed a different implementation of selective reconstruction based on binaural analysis called zero-crossing-based amplitude estimation (ZCAE) [36], illustrated in Fig. 3. The ZCAE algorithm estimates ITD in each frequency band by comparing zero crossings from the two microphones, which proved to be more effective than the more common method of estimating ITD from cross-correlation. A ratio mask that describes the probability that a given time-frequency segment of a sound is dominated by components from the target source is developed using analytical techniques. Figure 4 depicts word error rates (WERs) obtained for the DARPA Resource Management Corpus using ZCAE processing in connection with the CMU Sphinx-3 system. The environments used in the present study were digitally simulated using the RIR package [30] based on the image method, with the microphones placed 4 cm apart one another and 2 m from the speech sources. The target source was assumed to be arriving from a location along the perpendicular bisector of the line between the two microphones, while the masker is 45 degrees to one side. Results are compared for a single omnidirectional microphone (diamonds), the ZCAE algorithm making only a binary mask for each particular time-frequency segment (triangles), and the full ZCAE algorithm that implements a continuous estimate of the probability that the target dominates each segment using ratio masks (squares). The continuous ZCAE algorithm provides remarkably good performance in the absence of reverberation (Fig. 4), but its performance is severely degraded in the presence of reverberation, as shown in Fig. 5, which depicts performance at 10-dB SNR for various reverberation times.

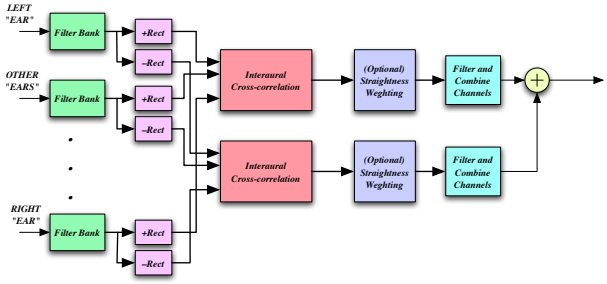

Fig. 6. *Schematic diagram of polyaural processing. Although only two microphones are depicted, the processing is extensible to an arbitrary number of microphones.*

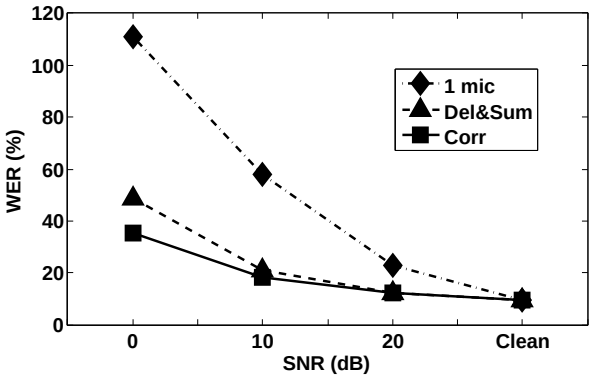

Fig. 7. *Speech recognition accuracy obtained for polyaural processing in the presence of an interfering speech source as a function of SNR in the absence of reverberation. Percentage WER is depicted for a single microphone (diamonds), delay-and-sum beamforming (triangles), and direct correlation processing (squares). See text for details.*

3.3. Correlation-based waveform enhancement

Most of the approaches discussed in the previous section involve processing from only two microphones. If more input channels are available, they can be exploited to reinforce the signal components from the target. The simplest such approach is traditional delay-and-sum forming (*e.g.* [37]) in which the inputs are subjected to digital delays which compensate for the acoustic path length differences of the target signal and the resulting phase-aligned target components are summed together. These approaches are relatively robust but they do not provide as much gain per input channel as other methods (*e.g.* [38]).

In an early study, Sullivan and Stern [39] elaborated on this approach by adding a correlation stage that crudely mimicked the peripheral auditory processing and inter-sensor correlation described in Sec. 2.4. The system provided substantial benefit in pilot experiments with artificially-combined signals but it was far less successful in real environments (for reasons we now realize are at least partially a consequence of the reverberation in those environments).

A more recent approach, which we refer to as polyaural processing elaborates on the approach of Sullivan and Stern modeling the cross-correlation function in more detail, again extending the

representation to more than two sensors (ears) [40]. Our model of the auditory periphery includes a bank of bandpass filters followed by half-wave rectification. Binaural (or polyaural) processing is modelled as a cross-correlation of the outputs of the filterbanks after rectification that are matched in terms of their best frequency of response. We also include an optional second level of cross-correlation that is performed across frequency, which serves to emphasize those components of the binaural response that are consistent over frequency or "straight" [22]. If there are only two sensors and only the positive portions of the filter outputs are considered, the processing described thus far is similar to that of the Jeffress-Colburn model. Waveform reconstruction is possible by adding a second (non-physiologically-based) half-wave rectifier that preserves only the negative portions of the filter outputs. These are combined across sensors in a similar fashion as the outputs of the positive rectifiers. After correlation across sensors (and possibly frequency) the outputs are normalized by taking the N^{th} root where N is the product of the number of sensors and frequency channels, passing the result through a bandpass filter similar to the initial filter (to remove the sharp "edges" caused by the halfwave rectification), and summed across frequency, as in Fig. 6. In principle, a signal arriving from the "look" direction in the absence of interference, reverberation, or noise will emerge from this processing without distortion, while components from other directions will be suppressed nonlinearly.

Figure 7 depicts word error rates (WERs) obtained for the DARPA Resource Management Corpus using polyaural processing in connection with the CMU SPHINX-3 system, using environmental conditions that were the same as the WER measurements for using the ZCAE algorithm described in Figs. 4 and 5. The input device presumed to be an 11-element logarithmic array of the type proposed by Flanagan [37] with three bands of 5-element subarrays with elements spaced at multiples of 4 cm. Results are compared for a single omnidirectional microphone (diamonds), a simple Flanagan delay-and-sum array (triangles), and the polyaural processing without weighting across frequency (squares). As is well known, array processing provides a dramatic improvement in WER compared to processing with a single microphone, even in the simple delay-and-sum configuration. The polyaural processing provides a relative improvement in WER of about 13.8 percent at 10 dB SNR and 27.8 percent at 0 dB compared to delay-and-sum beamforming. Stated another way, polyaural processing provides an effective improvement in SNR of roughly 3-5 dB at SNRs of 0 to 5 dB.

Figure 8 shows similar results in simulated reverberant environments with 2 m separating the talkers and the microphone array at an SNR of 10 dB. We note that delay-and-sum beamforming provides substantial improvement in these environments (which was not the case for the ZCAE algorithm), and that polyaural processing provides a further decrease in WER, resulting in an effective decrease in reverberation time of about 50 ms.

4. SUMMARY AND CONCLUSIONS

In this brief review we have described many of the binaural phenomena and models that have become the basis for computational processing intended to improve automatic speech recognition accuracy in cluttered and reverberant environments. Current speech processing systems have obtained impressive improvements in recognition accuracy in the absence of significant reverberation. The attainment of similar improvements in reverberant environments remains a serious challenge, and this is the major focus of current research efforts.

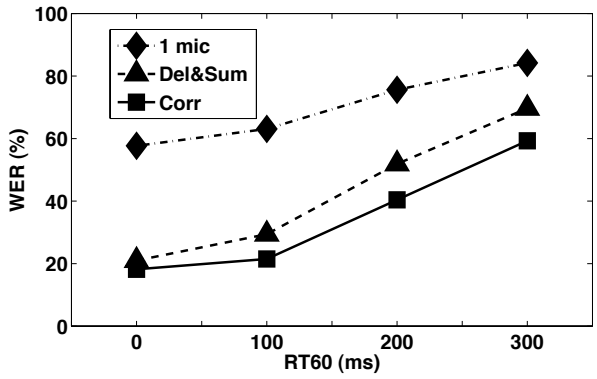

Fig. 8. *Similar to the previous figure, except that WER is measured in simulated reverberant environments with the source and array separated by 2 m. See text for details.*

5. REFERENCES

[1] J. W. Strutt (Lord Rayleigh), "On our perception of sound direction," *Philosoph. Mag.*, vol. 13, pp. 214–232, 1907.

[2] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform," *Journal of the Acoustical Society of America*, vol. 98, pp. 1980–1984, 1995.

[3] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 15, pp. 55–76, 1988.

[4] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenologial model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppresion," *Journal of the Acoustical Society of America*, vol. 109, pp. 648–670, 2001.

[5] A. R. Palmer, "Neural signal processing," in *Hearing*, B. C. J. Moore, Ed., Handbook of Perception and Cognition, chapter 3, pp. 75–121. Academic (New York), 1995.

[6] S. Kuwada, R. Batra, and D. C. Fitzpatrick, "Neural processing of binaural temporal cues," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., chapter 20, pp. 399–425. Lawrence Erlbaum Associates (Mahwah, New Jersey), 1997.

[7] J. E. Rose, N. B. Gross, C. D. Geisler, and J. E. Hind, "Some neural mechanisms in the inferior colliculus of the cat which may be relevant to localization of a sound source," *J. Neurophysiol.*, vol. 29, pp. 288–314, 1966.

[8] D. McAlpine, D. Jiang, and A. R. Palmer, "Interaural delay sensitivity and the classification of low best-frequency binaural responses in the inferior colliculus of the guinea pig," *Hearing Research*, vol. 97, pp. 136–152, 1996.

[9] N. I. Durlach and H. S. Colburn, "Binaural phenomena," in *Hearing*, E. C. Carterette and M. P. Friedman, Eds., vol. IV of *Handbook of Perception*, chapter 10, pp. 365–466. Academic Press, New York, 1978.

[10] R. Gilkey and T. R. Anderson, Eds., *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum, 1986.

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

[12] P. M. Zurek, "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, G. A. Studebaker and I. Hochberg, Eds. Allyn and Bacon, Boston, 1993.

[13] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *Journal of the Acoustical Society of America*, vol. 105, pp. 3436–3448, 1999.

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

[15] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psych.*, vol. 41, pp. 35–39, 1948.

[16] H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. I. general strategy and preliminary results on interaural discrimination," *Journal of the Acoustical Society of America*, vol. 54, pp. 1458–1470, 1973.

[17] R. M. Stern and H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. IV. a model for subjective lateral position," *Journal of the Acoustical Society of America*, vol. 64, pp. 127–140, 1978.

[18] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals," *Journal of the Acoustical Society of America*, vol. 80, pp. 1608–1622, 1986.

[19] J. Blauert, "Modeling of interaural time and intensity difference discrimination," in *Psychophysical, Physiological, and Behavioural Studies in Hearing*, G. van den Brink and F. Bilsen, Eds., pp. 412–424. Delft University Press, Delft, 1980.

[20] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. II. the law of the first wavefront," *Journal of the Acoustical Society of America*, vol. 80, pp. 1623–1630, 1986.

[21] W. Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," *Journal of the Acoustical Society of America*, vol. 94, pp. 98–110, 1993.

[22] 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., pp. 547–554. Pergamon Press, Oxford, 1992.

[23] 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*, 2003, vol. V, pp. 149–152.

[24] 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.

[25] G. J. Brown and K. J. Palomäki, "Reverberation," in *Computational Auditory Scene Analysis*, G. Brown and DeL. Wang, Eds. G. Brown and K. J. Palomäki, 2006.

[26] R. F. Lyon, "A computational model of binaural localization and separation," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1983, pp. 1148–1151.

[27] M. Bodden, "Modelling human sound-source localization and the cocktail party effect," *Acta Acustica*, vol. 1, pp. 43–55, 1993.

[28] M. Bodden and T. R. Anderson, "A binaural selectivity model for speech recognition," in *Proceedings of Eurospeech 1995*. European Speech Communication Association, 1995.

[29] 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.

[30] Stephen G. McGovern, *A model for room acoustics*, http://www.2pi.us/rir.html. 2004.

[31] D. R. Campbell, K. J. Palomaki, and G. J. Brown, "A matlab simulation of "shoebox" room acoustics for use in research and teaching," http://media.paisley.ac.uk/ campbell/Roomsim/.

[32] 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.

[33] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE Workshop on Applications of Signal Processing in Speech and Audio*. IEEE, 1997.

[34] N. Roman, S. Srinivasan, and DeL. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4040–4047, December 2006.

[35] S. Srinivasan, N. Roman, and DeL. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, pp. 1486–1501, 2006.

[36] H.-M. Park and R. M. Stern, "Spatial separation of speech signals using continuously-variable masks estimated from comparisons of zero crossings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[37] 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.

[38] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, P T R Prentice Hall, 1993.

[39] T. M. Sullivan and R. M. Stern, "Multi-microphone correlation-based processing for robust speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, 1993, pp. 91–94.

[40] R. M. Stern, E. B. Gouvêa, and G. Thattai, "Polyaural array processing for automatic speech recognition in degraded environments," in *Proc. of Interspeech 2007*. International Speech Communication Association, September 2007.