Binaural Technology for Machine Speech Recognition and Understanding

Richard M. Stern and Anjali Menon

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

Summary. It is well known that binaural processing is very useful for separating incoming sound sources as well as for improving speech intelligibility in reverberant environments. This chapter describes and compares a number of ways in which automatic speech recognition accuracy in difficult acoustical environments can be improved through the use of signal processing techniques that are motivated by our understanding of binaural perception and binaural technology. These approaches are all based on the exploitation of interaural differences in arrival time and intensity of the signals arriving at the two ears to separate signals according to direction of arrival and to enhance the desired target signal. Their structure is motivated by classic models of binaural hearing as well as the precedence effect. We describe the structure and operation of a number of methods that use two or more microphones to improve the accuracy of automatic speech recognition systems operating in cluttered, noisy, and reverberant environments. The individual implementations differ in the methods by which binaural principles are imposed on speech processing, and in the precise mechanism used to extract interaural time and intensity differences. Algorithms that exploit binaural information can provide substantially improved speech recognition accuracy in noisy, cluttered, and reverberant environments compared to baseline delay-and-sum beamforming. The type of signal manipulation that is most effective for improving performance in reverberation is different from what is most effective for ameliorating the effects of degradation caused by spatially-separated interfering sound sources.

1 Introduction

Automatic speech recognition (ASR) is the key technology that enables natural interaction between humans and intelligent machines. Core speech recognition technology developed over the past several decades in domains such as office dictation and interactive voice response systems to the point that it is now commonplace for customers to encounter automated speech-based intelligent agents that handle at least the initial part of a user query for airline flight information, technical support, ticketing services, etc. As time goes by, we will come to expect the range of natural human-machine dialog to grow to include seamless and productive interactions in contexts such as humanoid robotic butlers in our living rooms, information kiosks in large and reverberant public spaces, as well as intelligent agents in automobiles while traveling at highway speeds in the presence of multiple sources of noise. Nevertheless, this vision cannot be fulfilled until we are able to overcome the shortcomings of present speech recognition technology that are observed when speech is recorded at a distance from the speaker.

Two of the major forms of environmental degradation are additive noise of various forms and reverberation. Additive noise arises naturally from interfering speakers, background music, or other sound sources that are present in the environment, and as the signal-to-noise ratio (SNR) decreases, speech recognition becomes more difficult. In addition, the impact of noise on speech recognition accuracy depends as much on the type of noise source as on the SNR. For example, compensation becomes much more difficult when the noise is highly transient in nature, as is the case with many types of impulsive machine noise on factory floors and gunshots in military environments. Interference by sources such as background music or background speech is especially difficult to handle, as it is both highly transient in nature and easily confused with the desired speech signal. Research directed toward compensating for these problems has been in progress for more than three decades.

Reverberation is also a natural part of virtually all acoustical environments indoors, and it is a factor in many outdoor settings with reflective surfaces as well. The presence of even a relatively small amount of reverberation destroys the temporal structure of speech waveforms. This has a very adverse impact on the recognition accuracy that is obtained from speech systems that are deployed in public spaces, homes, and offices for virtually any application in which the user does not use a head-mounted microphone. It is presently more difficult to ameliorate the effects of common room reverberation than it has been to render speech systems robust to the effects of additive noise, even at fairly low SNRs. Researchers have begun to make meaningful progress on this problem only relatively recently.

In this chapter we discuss some of the ways fin which the characteristics of binaural processing have been exploited in recent years to separate and enhance speech signals, and specifically 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 as of yet we do not know how best to achieve this level of performance using the engineering tools available in contemporary signal processing.

In the next section we restate very briefly the basic binaural phenomena that have been exploited in contemporary signal enhancement and robustness algorithms for ASR. In Sec. 3 we summarize for the lay person some of the basic principles that underly contemporary ASR systems. We survey a number of computational approaches to impove the accuracy of ASR systems that are motivated by binaural processing in Sec. 4 and we discuss some extensions of these approaches to systems based on deep learning in Sec. 5.

2 Binaural-hearing Principles

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. These capabilities have motivated a great number of studies of binaural physiology and perception. Useful comprehensive reviews of basic binaural perceptual phenomena may be found in a number of sources including Durlach and Colburn (1978), Gilkey and Anderson (1997), Stern *et al.* (2006), and Kolrausch *et al.* (2013), among others, as well as in basic texts on hearing such as Moore (2012) and Yost (2013).

2.1 Selected Binaural Phenomena

While the literature on binaural processing on both the physiological and perceptual sides is vast, the application of binaural processing to ASR is based on a small number of principles:

1. The perceived laterality of sound sources depends on both the interaural time difference (ITD) and interaural intensity difference (IID) of the signals arriving to the two ears, although the relative salience of these cues depends on frequency (*e.g.* Durlach and Colburn, 1978; Domnitz and Colburn, 1977; Yost, 1981).
2. The auditory system is exquisitely sensitive to small changes of 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 broadband noise sources is also quite acute, as a decrease in interaural correlation from 1.00 to 0.96 is readily discernible (*e.g.* Durlach and Colburn, 1978; Domnitz and Colburn, 1977). The ITDs arise from differences in path length from a sound source to the two ears, and the IIDs are a consequence of head shadowing, especially at higher frequencies.
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 anatomy of the outer ear, and reinforced by head-motion cues (*e.g.* Mehrgardt and Mellert, 1977; Wightman and Kistler, 1989a,b, 1999). The transfer function from the sound source to the ears is commonly referred to as the *head-related transfer function* (HRTF). HRTFs generally depend on the azimuth and elevation of the source relative to the head, as well as the anatomy of the head and outer ear of the individual.

4. The intelligibility of speech in the presence of background noise or some other interfering signal becomes greater as the spatial separation between the target and masking signals increases. While some of the improvement in intelligibility with greater spatial source separation may be attributed to monaural effects such as a greater effective SNR at one of the two ears, binaural interaction also appears to play a significant role (*e.g.* Zurek, 1993; Hawley *et al.*, 1999).
5. The auditory localization mechanisms typically pay greater attention to the first component that arrives (which presumably comes directly from the sound source) at the expense of later-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* (*e.g.* Wallach *et al.*, 1949; Blauert, 1997; Litovsky *et al.*, 1999).

2.2 Models of Binaural Interaction

A number of models have been developed that attempt to identify and explain the mechanisms that mediate the many interesting binaural phenomena that have been observed. For the most part, the original goals of these models had been to describe and predict binaural lateralization or localization, discrimination, and detection data, rather than to improve ASR recognition accuracy. These models are typically evaluated on their ability to describe and predict the perceptual data, the generality of their predictions, and the inherent plausibility of the models in terms of what is known about the relevant physiology. Useful reviews of binaural models may be found in Colburn and Durlach (1978), Stern and Trahiotis (1995, 1996), Trahiotis *et al.* (2005), Braasch (2005), Colburn and Kulkarni (2005), and Dietz *et al.* (2017), among other sources.

Most theories of binaural interaction (at least for signals that are presented through headphones) include a model that describes the peripheral response to sound at the level of the fibers of the auditory nerve, a mechanism for extracting ITDs, a mechanism for extracting IIDs, a method for combining the ITDs and IIDs, and a mechanism for developing predictions of lateral position from the combined representation. Models that describe sound localization in the free field typically incorporate information from HRTFs.

Models of Auditory-nerve Activity

Models of the response to the sounds at the auditory-nerve level typically include (1) a bandpass frequency response, with a characteristic frequency (CF) that provides the greatest response, (2) some sort of half-wave rectification that converts the output of the bandpass linear filters to a strictly positive number that represents rate of response, and (3) synchrony or “phase locking” in the response to the fine structure of low-frequency inputs and to the envelopes of higher-frequency inputs. Some auditory-nerve models also include

(4) enhanced response at the temporal onset of the input and (less frequently) (5) an explicit mechanism for lateral suppression in fibers with a given CF to signal components at adjacent frequencies. These models of auditory-nerve activity can be as simple as the cascade of a bank of bandpass filters, half-wave rectification, and lowpass filtering; more complex and physiologically-accurate models are described in Zhang *et al.* (2001) and Zilany *et al.* (2009), among other sources.

Cross-Correlation-Based Models

Most models of binaural interaction include some form of Jeffress's (1948) description of a neural "place" mechanism as the basis for the extraction of interaural timing information. Specifically, 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. It was 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. Because of the synchrony in the response of low-frequency auditory-nerve fibers to low-frequency signals, 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 (1969, 1973) reformulated Jeffress's hypothesis quantitatively using a relatively simple model of the auditory-nerve response to sound as Poisson processes, 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. The overall response of an ensemble of such units as a function of internal delay is similar to the running interaural cross-correlation of the signals to the two ears, after the peripheral cochlear analysis (*e.g.* Stern and Trahiotis, 1995). This general representation has been used in a number of computational models of binaural processing for speech recognition, with sound-source locations identified by peaks of the interaural cross-correlation functions along the internal-delay axis.

Figure 1 illustrates how the Jeffress-Colburn mechanism can be used to localize two signals according to ITD. The upper two panels of the figure show the magnitude spectra in decibels of the vowels /AH/ and /IH/ spoken by a male and a female speaker, respectively. The lower panel shows the relative response of the binaural coincidence-counting units when these two vowels are presented simultaneously with ITDs of 0 and -0.5 ms, respectively. The 700-Hz first formant of the vowel /AH/ is clearly visible at the 0-ms internal delay, and the 300-Hz first formant of the vowel /IH/ is seen at the delay of -0.5 ms.

It should be noted that the interaural cross-correlation function does not describe IIDs unambiguously, so some additional mechanism must be em-

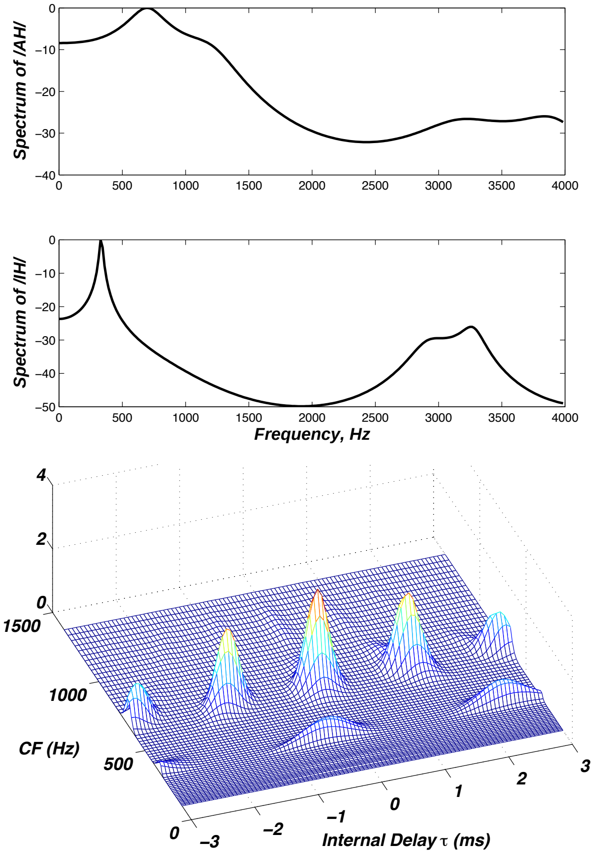

Fig. 1. Upper and central panels: spectrum of the vowels /AH/ and /IH/ as recorded by a male and female speaker, respectively. Lower panel: response of an implementation of the Jeffress-Colburn model to the simultaneous presentation of the /AH/ presented with a 0-ms ITD and the /IH/ presented with a −0.5-ms ITD.

ployed to represent the contributions of IID. For example, Stern and Colburn (1978) multiplied the cross-correlation-based representation of ITD described above by a pulse-shaped function with a location along the internal-delay axis that depends on IID. This model, known as the "position-variable model," predicts lateral position by computing the centroid of the product of these "timing" and "intensity" functions along the internal-delay axis and then integrating this function over characteristic frequency. Shamma *et al.* (1989) proposed an alternative implementation of the Jeffress model, called *stereausis* in which the internal delays are obtained implicitly by comparing inputs of auditory-nerve fibers with slightly mismatched characteristic frequencies, as previously suggested by Schroeder (1977).

Blauert and his colleagues proposed a similar representation (Blauert and Cobben, 1978; Blauert, 1980). This work was subsequently extended by Lindemann (1986a), who added a mechanism that (among other things) inhibits outputs of the coincidence counters when there is activity produced by coincidence counters at adjacent internal delays. This contralateral inhibition mechanism enables the Lindemann model to describe several interesting phenomena related to the precedence effect (Lindemann, 1986b). Gaik (1993) extended the Lindemann mechanism further by adding a second weighting to the coincidence-counter outputs that reinforces naturally-occurring combinations of ITD and IID.

The Equalization-Cancellation model

The Equalization-Cancellation (EC) model of Durlach and colleagues (*e.g.* Durlach, 1963, 1972) is an additional important alternate model. The EC model was initially formulated to account for binaural detection phenomena, although it has been applied to other psychoacoustical tasks as well (Colburn and Durlach, 1978). The model assumes that time-delay and amplitude-shift transformations are applied to the incoming signal on one side in order to *equalize* the masker components of the signals to the two ears. The masker-equalized signals are then subtracted from one another to *cancel* the masker components, leaving the target easily detectable. Stochastic "jitter factors" are applied to the time and amplitude transformations, which limits the completeness of the equalization and cancellation operations, in a fashion that is fitted to the observed limits of human detection performance. The EC model remains popular because of its simplicity and its ability to describe many phenomena. It has been the inspiration for subsequent models (*e.g.* Breebaart *et al.*, 2001a,b,c), and has also been applied to speech recognition, as will be discussed below.

Detection of Target Presence Using Interaural Correlation

Many phenomena, especially in the area of binaural detection, can be interpreted easily by considering the change in interaural correlation that occurs when a target is added to the masker. The use of interaural correlation was formalized in one binaural early model (Osman, 1971) and has been the focus of many experimental and theoretical studies since that time, as reviewed by Trahiotis *et al.* (2005) among other sources. While cross-correlation-based models that represent ITD, the EC model, and correlation-based models differ in surface structure, it has been shown that under many circumstances they function similarly for practical purposes (*e.g.,* Colburn and Durlach, 1978; Domnitz and Colburn, 1976).

3 Selected Robust Speech-recognition Principles

The field of robust automatic speech recognition is similarly vast, and cannot be dealt with in any depth in a review chapter of this scope. The purpose of this section is to provide some insight into the principles of automatic speech recognition that are needed to appreciate the role that binaural processing can play in reducing error rates.

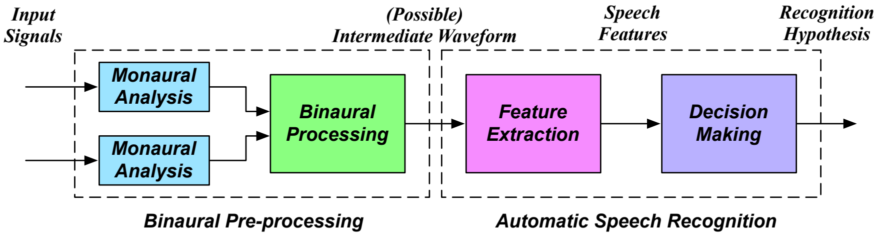

Fig. 2. Basic functional elements of a speech recognition system that includes binaural enhancement.

3.1 Basic Speech-recognition Principles

Automatic speech recognition is essentially a special class of *pattern classification* algorithms, that guess which of a number of possible "classes" of input is actually present. All pattern classification systems operate on the same basic principles: an initial analysis stage performs a physical measurement (of a sound pressure wave, in our case) and transforms that measurement into a set of *features*, or numbers that are believed to be most indicative of the classification task to be performed. These features are typically a stochastic representation that depends on which input class is present. A second decision-making component develops a hypothesis of which of the possible inputs is most likely, based on the observed values of the features. Figure 2 summarizes the major functional blocks of a generic ASR system with binaural pre-processing for signal enhancement. While Fig. 2 depicts a binaural pre-processing module that passes on to the ASR components a restored speech waveform, some of the algorithms we describe produce a restored set of features directly. We briefly discuss the components of the speech recognition system in this section and defer our discussion of the numerous approaches to signal and feature enhancement based on binaural processing to Sec. 4 below.

Feature Extraction

Features for pattern classification systems are generally selected with the goals of being useful in distinguishing the classes to be identified, easy to compute, and not very demanding in storage. With some exceptions, most speech recognition systems today extract features by first computing the *short-time Fourier transform* (STFT) of the input signal (Allen and Rabiner, 1977), typically windowing the incoming signal by a succession of Hamming windows of duration approximately 25 ms, separated by approximately 10 ms. A function related to the log of the magnitude of the spectrum or its inverse transform, the *real cepstrum*, is subsequently computed in each of these analysis frames. In principle, cepstral coefficients are useful because they are nearly statistically independent of one another, and only a small number of them (about 12) are needed to characterize the envelope of the spectrum in each analysis frame. In addition, the cepstral representation separates the effects of the vocal-tract filter (which were believed to be most useful in the early days of the speech recognition) from the effects of the periodic excitation produced by the vocal cords (which had been believed not to be useful at that time).

The most common representations used for feature vectors today are all motivated by crude models of auditory processing. The earliest such representation, *mel frequency cepstral coefficients* (MFCC features, Davis and Mermelstein, 1980), multiply the energy spectrum extracted from each analysis frame by a series of triangularly-shaped weighting functions with vertices spaced according to the Mel frequency scale (Stevens *et al.*, 1937) and then summing the product over frequency within each weighting function. With 16-kHz sampling, about 40 Mel weighting functions are typically used. The MFCC coefficients are obtained by computing the inverse discrete cosine transform (DCT) of the summed products. A second set of popular features are extracted using a process known as *perceptual linear prediction* (PLP features, Hermansky, 1990), which is based on a more detailed and accurate model of the peripheral auditory system. A more recently-developed third set of features, *power-normalized cepstral coefficients* (PNCC features, Kim and Stern, 2016) are more robust to certain types of additive noise and reverberation.

The MFCC, PLP, or PNCC features are typically augmented by additional features that represent the instantaneous power in each analysis frame, as "delta" and "delta-delta" features that serve to represent crudely the first and second derivatives in the power spectrum over time. The delta features are obtained by computing the difference between cepstral coefficients in frames after and before the nominal analysis frame, and the delta-delta features are obtained by repeating this operation. Finally, static effects of linear filtering to the signal are removed by applying either *cepstral mean normalization* (CMN), or *relative spectral analysis* (RASTA) processing (Hermansky and Morgan, 1994). CMN subtracts the mean of the cepstral coefficients from each cepstral vector on a sentence-by-sentence basis while RASTA processing passes the cepstral coefficients through a bandpass filter. Both RASTA and CMN serve to emphasize temporal change in the cepstral coefficients and suppress slow drift in their values over time.

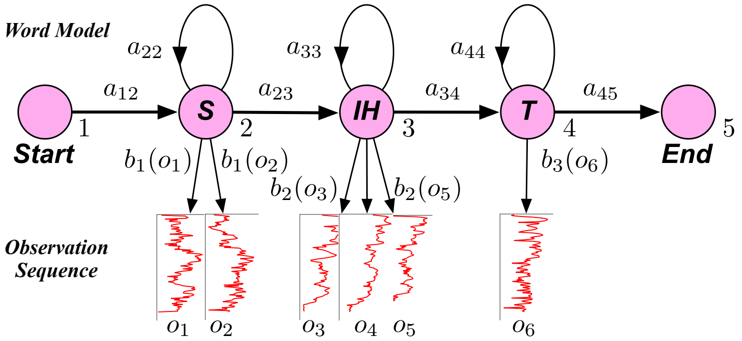

Fig. 3. The hidden Markov model for speech recognition for the word "sit." See text for details.

Traditional HMM-GMM Decoding

The technologies for determining the most likely word sequence from a spoken utterance have evolved greatly over the decades, and this section will discuss only the most basic elements of speech recognition. From the early 1980s until very recently the dominant speech recognition technology has been the *hidden Markov model* (HMM, *e.g.,* Rabiner, 1989; Rabiner and Juang, 1993), and practical systems based on HMMs remain in widespread use today. The HMM representation characterizes the incoming speech waveform as a doubly stochastic process, as depicted in schematic form in Fig. 3. First, the sequence of phonemes that are produced is characterized as a set of five unobserved Markov states which presumably represent the various configurations that the speech production mechanisms may take on and hence the phonemes that are produced. As is the case for all Markov models, the transition probabilities depend only on the current state that is being occupied. Each state transition causes a feature vector to be emitted that is observable, with the probability density of the components of the feature vector depending on the identity of the state transition. Spectra representing a sequence of six observations are shown in the figure. The task of the decoder is to infer the identity of the unobserved state transitions (and hence the sequence of phonemes) from the observed values of the features.

The technologies for implementing this model efficiently and accurately have evolved greatly over decades, and a detailed description is well beyond the scope of this chapter. Briefly, implementing an HMM requires determining the probabilities of the observations given the model parameters, choosing the most likely state sequence given the observations, and determining the model parameters that maximizes the observation probabilities. Details of how to accomplish these tasks are described in standard texts such as Rabiner and Juang (1993) and Gold *et al.* (2011), as well as in many technical papers. It has been found that the performance of the system depends more critically on the accuracy of the phonetic model (*i.e.*, the probability density function that describes the feature values given the state transitions) than on the probabilities that characterize the state transitions. Gaussian mixture densities are currently the form that is most commonly used for the phonetic models, in part because the parameters of these densities can be estimated efficiently, typically using a form of the expectation-maximation (EM) algorithm (*e.g.,* Dempster *et al.*, 1977). HMMs using Gaussian mixtures for the phonetic models are frequently referred to as "HMM-GMM" systems.

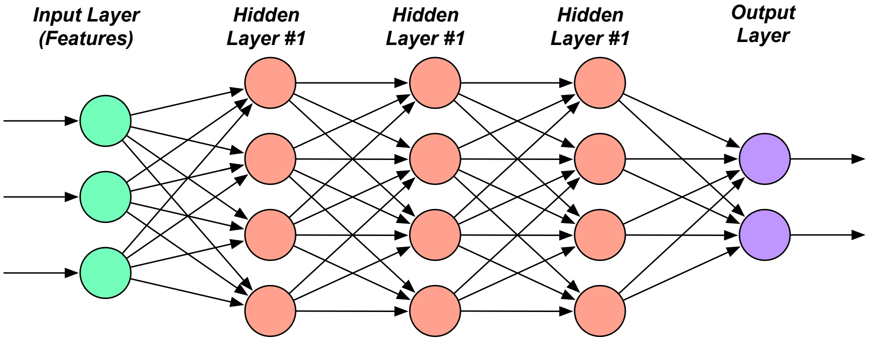

Fig. 4. Standard structure of a feedforward MLP. The network is considered to be "deep" if there are two or more hidden layers.

Speech Recognition Using Deep Learning

While the HMM-GMM paradigm has been the dominant speech recognition technology from the early 1980s through the mid-2000s, new approaches to speech recognition based on *deep learning* are becoming more popular. The structures that implement deep learning are frequently referred to as *artificial neural networks* or *computational neural networks*. The general organization and function of computational neural networks was originally motivated by basic neural anatomy and physiology, although the classifiers have evolved considerably over the years without any necessary tie to neural processing by living beings.

While the basic approaches to pattern classification using computational neural networks have been known for some time (*e.g.,* Rosenblatt, 1959; Lippmann, 1987, 1989; Bourlard and Morgan, 1994), these approaches have become more effective and practical in recent years because of a better understanding of the capabilities of the underlying mathematics, the widespread availability of much larger databases for training, and much faster computing infrastructure, including the availability of *graphics processing units* (GPUs), which are particularly well suited for many of the core computations associated with neural networks.

Figure 4 is a crude depiction of the simplest type of deep neural network (DNN) known as the *multilayer perceptron* (MLP). The system consists of an input layer of units, one or more "hidden layers," and an output layer. Typically the units in a given layer are a weighted linear combination of the values of the units of the previous layer, with the values of the weights trained to minimize the mean square error of the result, using a technique based on gradient descent known as *back propagation* (*e.g.,* Haykin, 2018). In many cases DNN classifiers make use of observed values of multiple feature sets (*e.g.,* Mitra *et al.*, 2017). In general, computational neural networks have the advantage of being able to model probability density functions of any form by learning their shape by observing large numbers of training examples. They have the disadvantage of requiring more training data than conventional HMM-GMM systems, and they may not generalize as well as HMM-GMM-based systems. While neural networks were initially used to produce better phonetic models in a system that incorporated a traditional HMM for the decoding component (*e.g.* Hermansky *et al.*, 2000), other architectures are becoming more popular in which the entire end-to-end speech recognition process is performed using a chain of deep neural networks (*e.g.* Miao and Metze, 2017). Nevertheless, they are increasingly popular because they provide consistently better acoustic-phonetic models than the traditional Gaussian mixtures. The technologies of deep learning have undergone explosive growth and development in recent years, and the reader is referred to standard texts and tutorials such as Goodfellow *et al.* (2016) and Nielsen (2016) for detailed explanations of the technology.

3.2 Signal Processing for Improved Robustness in ASR

We discuss briefly in this section some of the traditional approaches that have been applied to signals to improve recognition accuracy in ASR systems. This field is vast, and has been the object of very active research for decades. Excellent recent reviews of a variety of techniques may be found in Virtanen *et al.* (2012). In this section we focus on basic feature enhancement techniques, missing-feature approaches, and the uses of multiple microphones.

Feature-based Compensation for Noise and Filtering

Many successful approaches to robustness in ASR are direct descendants of approaches that were first proposed to enhance speech for human listeners. For example, *spectral subtraction* (Boll, 1979), reduces the effects of additive noise by estimating the magnitude of the noise spectrum and subtracting it on a frame-by-frame basis from the spectrum of the signal, reconstructing the time-domain signal with the original unmodified phase. This approach was the basis of dozens if not hundreds of subsequent noise-mitigation algorithms. Stockham *et al.* (1975) proposed the use of *homomorphic deconvolution* to mitigate the effects of linear filtering by, in effect, subtracting the log magnitude spectrum (or its inverse transform, the real cepstrum) of an estimate of the sample response of the unknown linear filter. A simplified version of this approach is the basis for the cepstral mean normalization that is widely used in ASR systems today.

Joint compensation for the effects of noise and filtering is complicated by the fact that they combine nonlinearly: noise is additive in the time and frequency domains while the effects of filtering are additive in the log spectral and cepstral domains. One particularly successful approach has been the *vector Taylor series* (VTS) algorithm (Moreno *et al.*, 1996), which models the degraded speech as clean speech passed through an unknown linear filter and subjected to unknown additive noise. The algorithm estimates the parameter values that characterize the filtering and noise in a fashion that maximizes the probability of the observations. A recent review of VTS and a number of other techniques motivated by it may be found in Droppo (2013). Algorithms like VTS can provide good improvements to recognition accuracy when the statistics characterizing the noise and filtering are quasi-stationary while parameters are being estimated, but they are less effective when disturbances are more transitory as in the case of background music or a single interfering speaker. The use of missing-feature approaches as described below has been more effective for these signals.

Computational Auditory Scene Analysis and Missing-feature Approaches

Modern missing-feature approaches to robust recognition are inspired by Bregman's seminal work (Bregman, 1990) in *auditory scene analysis*. Bregman examined the cues that people appear to use in order to segregate and cluster the various components that belong to individual sound sources while perceiving multiple sources that are presented simultaneously. Cues that have proved to be useful include commonalities in onset, amplitude modulation, frequency modulation, and source location, along with harmonicity of components, among others.

Computational auditory scene analysis (CASA) refers to a number of approaches that attempt to emulate the perceptual segregation of sound sources using computational techniques (*e.g.* Brown and Cooke, 1994; Cooke and Ellis, 2001; Wang and Brown, 2006). The implementation of CASA to isolate the desired signal for an ASR system typically begins by determining which components of the incoming signal are dominated by the target signal and hence not distorted or "missing." In ASR systems, the initial representation is typically in the form of a spectro-temporal display such as a spectrogram. Consideration of only those elements that are relevant or undistorted can be thought of as a multiplication of the components of the spectrogram by a "binary mask" (if "yes-no" decisions are made concerning the validity of a particular spectro-temporal component) or by a "ratio mask" (if probabilistic decisions are made). Once a mask is developed, speech recognition is performed by considering only the subset of components that are considered to be "present" (*e.g.,* Cooke *et al.*, 2001), or by inferring the values of the "missing" features (*e.g.,* Raj *et al.*, 2004) and performing recognition using the reconstructed feature set.

While signal separation and subsequent ASR using CASA techniques can be quite effective if the binary or ratio mask is estimated correctly, (*e.g.* Cooke *et al.*, 2001; Raj *et al.*, 2004; Raj and Stern, 2005), estimating the mask correctly is frequently quite difficult in practice, especially when little is known *a priori* about the nature of the target speech and the various sources of degradation. One singular exception to this difficulty in estimating the masks correctly arises when signals are separated in space and the target location is known, as components can be relatively easily separated using ITD-based and IID-based information. For this reason, separation strategies motivated by binaural hearing have been quite popular over the years for speech recognition systems that make use of two microphones.

Figure 5 shows sample spectrograms of signals separated according to ITD in anechoic and reverberant rooms using two microphones. The speech sources were placed 2 m from the microphones, and at an angle of ±30 degrees from the perpendicular bisector of a line connecting the microphones. The microphones were 4 cm apart and the room impulse response (RIR) simulation package McGovern (2004) was used to develop the simulated impulse responses of the room. The rows of the figure depict, in order, spectrograms of the left speech source, the right speech source, the two sources combined, the separated left source, and the separated right source. By comparing the spectrograms in rows (a) and (d), and (b) and (e), it can be seen that the separation is much more effective when the speech is not reverberated.

Conventional Signal Processing using Multiple Microphones

The benefit provided by any approach that attempts to improve ASR accuracy using binaural approaches must be compared to the improvement produced by a similar configuration of microphones using conventional techniques. These conventional approaches, frequently referred to as *beamforming algorithms*, attempt to develop a response that is most sensitive to signals coming from a particular "look direction" while either being less sensitive to sources from other directions or actively nulling the responses to these other sources. Classical multi-microphone signal processing techniques are highly developed and discussed in texts including Johnson and Dudgeon (1993) and Van Trees (2004). Recent results concerning the application of multi-microphone techniques to ASR are summarized in Kumatani *et al.* (2012).

The simplest multi-microphone technique is *delay-and-sum beamforming* in which the path length differences from the target source to the various microphones are compensated for by time delays imposed by the system to ensure that the target signal components from the various microphones always arrives at the same time to the system, creating constructive interference. Signal components from other directions will combine constructively or destructively

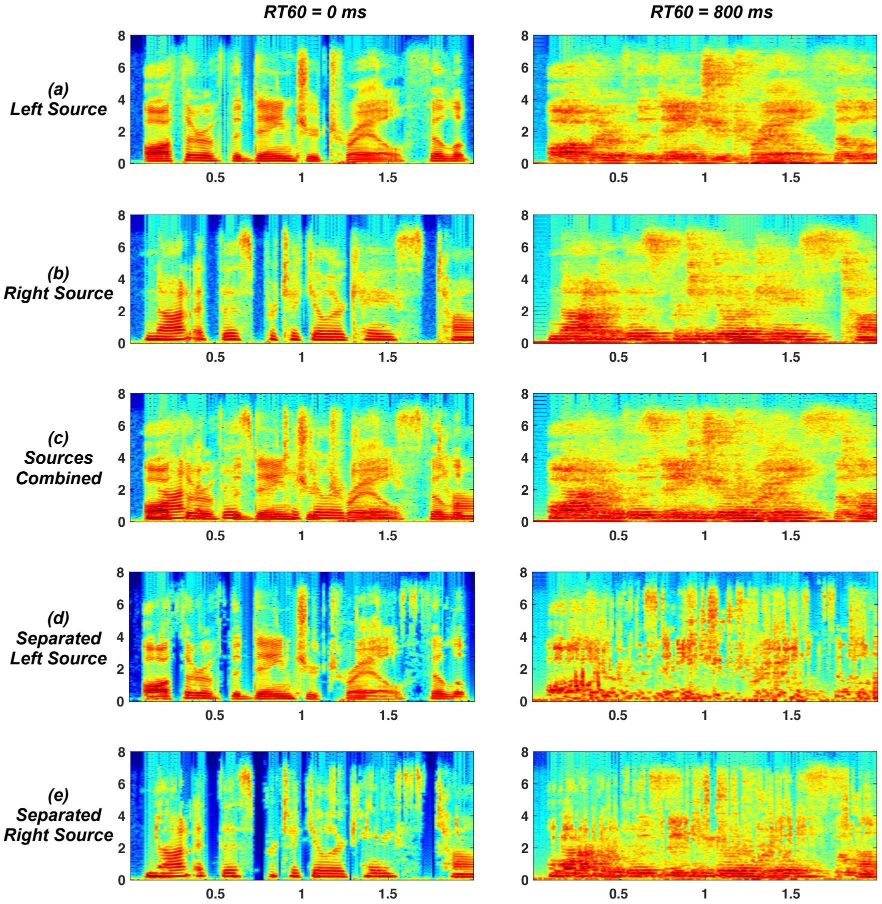

Fig. 5. Sample spectrograms of two speech signals separated according to ITD. The original signals in the left column are clean speech, while the signals in the right column were convolved with a simulated room impulse response with a reverberation time of 800 ms. The spectrograms represent (a) the signal on the left side, (b) the signal on the right side, (c) the two signals combined, (d) the signal on the left side separated from the combined signal according to ITD, (e) the separated signal on the right side. The horizontal axis is time in s and the vertical axis is frequency in kHz

across the microphones, and hence they would be reinforced to a lesser degree, on average. Because the actual directional sensitivity depends on an interaction between the wavelength at a given frequency, the directivity pattern for delay-and-sum beamforming varies with frequency. Generally the width of the main lobe decreases as frequency increases, and eventually "spatial aliasing" will occur when an interfering signal component arrives at a frequency and azimuth such that the distance between the microphones becomes greater than half a wavelength. These frequency effects can be mitigated by the use of nested arrays with different element spacings (*e.g.* Flanagan *et al.*, 1985) and by the use of *filter-and-sum* beamforming techniques in which the fixed delays in delay-and-sum beamforming are replaced by discrete-time linear filters which can in principle impose different delays at different frequencies for each microphone.

Modern techniques such as the *minimum variance distortionless response* (MVDR) method use minimum mean square estimation (MMSE) techniques that seek to maintain a fixed frequency response in the look direction while at the same time suppressing the response from the directions of arrival of the most powerful interfering sources (*e.g.* Van Trees, 2004). The performance of these optimum linear signal processing approaches to multi-microphone beamforming also degrades in reverberant environments because the phase incoherence imposed by the reverberance causes the estimation of important statistics such as the auto- and cross-correlations of the signals across the microphones to become much less accurate. McDonough and others have achieved some success with the use of objective functions based on negative entropy or kurtosis as the basis for optimizing the filter coefficient values (*e.g.* Kumatani *et al.*, 2012). These statistics drive the coefficients of the arrays to produce output amplitude histograms that are “heavier” in the tails, which corresponds to output that is more speech-like than the Gaussian densities that characterize sums of multiple noise sources.

4 Binaural Technology in Automatic Speech Recognition

In this section we describe and discuss selected methods by which ASR accuracy can be improved by signal processing approaches that are motivated by binaural processing. Most of the systems considered improve ASR accuracy by some sort of selective reconstruction of the target signal using CASA-motivated techniques, which use differing approaches to identify the subset of spectro-temporal components in the input that are dominated to the greatest extent by the target signal. The most common approach makes this determination by comparing measured ITDs and IIDs for each spectro-temporal component to the values of these parameters that would be observed from a source arriving from the putative target direction, as described below in Sec. 4.2. A second approach is based on the value of the overall normalized interaural cross-correlation, as spectro-temporal components with high interaural cross-correlation are more likely to be dominated by a single coherent target signal, as described in Sec. 4.4. A third approach implements a modification to the EC model, in which the two inputs are equalized according to the nature of the *target* signal, and then subtracted from one another, as described in Sec. 4.5. This causes the spectro-temporal components that are dominated the most by the target signal to change by the greatest amount.

In addition to the three methods above used to identify the most relevant spectro-temporal components of the input, the systems proposed also differ in other ways including the following:

- The extent to which a particular system is intended to provide a complete auditory scene analysis, including identification, localization, and classification of multiple sources versus simply providing useful enhancement of a degraded primary target signal for improved speech recognition accuracy.
- Whether the location of the desired target is expected to be estimated by the system or is simply assumed to be known *a priori*.
- Whether a particular system is designed to receive its input from two ears on a human or manikin head rather than two (or more) microphones in the free field. The use of a real or simulated head provides IIDs and the opportunity to use them to disambiguate the information provided by ITD analysis. In contrast, systems that do not include an artificial head are typically easier to implement, and the absence of a head facilitates the use of more than two microphones.
- Whether a particular system works by reconstructing a continuous-time enhanced speech waveform that is processed by the normal front end of an ASR system or whether it simply produces enhanced features representing the input such as cepstral coefficients and inputs these enhanced features directly into the ASR system.
- The nature of the acoustical environment, including the presence or absence of diffuse background noise, coherent interfering sound sources, and/or reverberation, etc. within which a particular system is designed to operate.

It is worth noting that researchers at the University of Sheffield and Ohio State University, working in collaboration or independently, have provided the greatest number of contributions to this field over the years, both in terms of fundamental principles and system development. Interesting contributions over many years have also been provided by groups at the universities at Bochum and Oldenburg in Germany, as well as a number of other locations around the world including our own university.

The representative systems considered do not sort themselves into convenient mutually-exclusive categories, so we somewhat arbitrarily have sorted our discussion according to how the most relevant spectro-temporal targets components are identified, as discussed above. We begin with a brief summary of some of the earliest attempts to apply binaural processing to improve ASR accuracy. We then summarize the organization of representative systems based on extraction of ITD and IID information using CASA principles. We continue with a discussion of the use of onset enhancement to ameliorate the effects of reverberation, the development of systems based on interaural coherence, and approaches based on the EC model. We conclude with some brief comments on various approaches used to extend these approaches to more than two incoming signals.

4.1 Early Approaches

Lyon (1984) proposed one of the first systems applying binaural hearing principles, using a computational model of auditory-nerve activity from two sources as an input to a Jeffress-like network of coincidence-counting units. He suggested that this structure could be applied to multiple applications including ASR. While Lyon's system was not evaluated quantitatively because the ASR systems of the day were mathematically primitive and computationally costly, he noted that this approach appeared to provide a stable spectral representation for vowels as well as source separation according to ITD.

Most evaluations of ASR with binaural processing in the early period consisted of the concatenation of an existing binaural model with a speech recognition system. For example, Bodden (1993) described an early CASA-based system, called the Cocktail-Party-Processor (CPP) that had many of the elements of later systems, implementing a structure suggested by Blauert (1980). The CPP included HRTFs that introduced frequency-dependent ITDs and IIDs based on angle of arrival, a relatively simple auditory-nerve model that included bandpass filtering, half-wave rectification, lowpass filtering, and saturation of the rate of response. Binaural processing in the CPP incorporated the Lindemann (1986a) model with contralateral inhibition, which predicted certain precedence-effect phenomena and appropriate intereactions between ITD and IID, and the additional contributions of Gaik (1993), which developed lateralization information from ITDs and IIDs in a fashion that was cognizant of the natural combinations of these interaural differences as observed in HRTFs. In later work, Bodden and Anderson (1995) used a simple speech recognizer, the self-organizing feature map (SOFM) of Kohonen (1989), and demonstrated improved ASR accuracy for simple phonemes in the presence of spatially-separated noise, especially at lower SNRs. DeSimio *et al.* (1996) obtained similar results with a different auditory-nerve model (Kates, 1991) and Shamma's stereausis model (Shamma *et al.*, 1989) to characterize the binaural interaction.

4.2 Systems Based on Direct Extraction of ITD and IID Information

By far the most common application of binaural principles to ASR is through systems that implement computational auditory scene analysis using direct extraction of ITDs and IIDs in some fashion, as depicted in Fig. 6. In general, these systems attempt to estimate the extent to which each spectro-temporal component of the input is dominated by the target signal based on ITDs and IIDs that are extracted. We summarize in this section a few of the methods that are used to implement each component in representative systems.

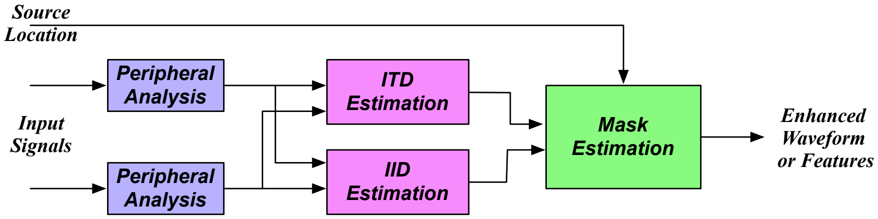

Fig. 6. Functional blocks of a CASA-based system that extracts ITDs and IIDs directly. The input may be from free-field microphones or through a real or simulated head. The source location may be known or estimated.

Extraction of Natural Interaural differences

As noted above, a number of the systems develop their hypotheses from naturally-extracted interaural differences (*e.g.,* Roman *et al.*, 2003; Palomäki *et al.*, 2004; Srinivasan *et al.*, 2006; Brown *et al.*, 2006; Harding *et al.*, 2006; May *et al.*, 2011, 2012). While these systems typically made use of measured HRTFs (*e.g.,* Gardner and Martin, 1994) obtained through the use of the KEMAR manikin (Burkhard and Sachs, 1975), they could also have been obtained in principle using small microphones in the ear canals (*e.g.,* Wightman and Kistler, 1989a). Because the relationship between ITD and the azimuth of the source location in HRTFs depends weakly on frequency, some systems (*e.g.,* Roman *et al.*, 2003; Palomäki *et al.*, 2004; May *et al.*, 2011) incorporated an explicit mapping table that converts ITD into putative arrival angle in a manner that is consistent across all frequencies. The IIDs show significant dependencies on both azimuth and frequency. For the most part, the various sound sources were assumed to be at the same elevation as the microphones.

Other systems (*e.g.,* Aarabi and Shi, 2004; Park and Stern, 2009; Kim *et al.*, 2009) work from free-field input without an artificial head or HRTFs, and consequently the masks that are produced cannot make use of IID information.

Peripheral Auditory Processing

All binaural processing systems incorporate some abstraction of the frequency-dependent processing imparted by the peripheral auditory system. The most common approach (*e.g.,* Roman *et al.*, 2003; Palomäki *et al.*, 2004; Harding *et al.*, 2006) is to use a bank of 40 to 128 Gammatone filters (Patterson *et al.*, 1988), followed by half-wave rectification, lowpass filtering (which provides envelope extraction at higher frequencies), and in some cases nonlinear compression of the resulting signal. Other systems (*e.g.,* Kim *et al.*, 2009) simply compute the short-time Fourier transforms (STFTs) of the two input signals, from which ITDs and IIDs can be inferred by comparison of the magnitudes and phases for each spectro-temporal component.

Estimation of ITDs and IIDs

There are multiple ways of extracting ITDs from the results of the peripheral processing. The most common approach is to compute a variant of the *normalized interaural cross-correlation function* at each frequency:

$$R[m,k] = \frac{\sum_{n=0}^{N-1} x_{L,k}[n] x_{R,k}[n-m]}{\sqrt{\sum_{n=0}^{N-1} x_{L,k}^2[n]}\sqrt{\sum_{n=0}^{N-1} x_{R,k}^2[n-m]}}$$

where $R[m,k]$ is the normalized interaural cross-correlation as a function of lag m and frequency index k, and $x_{L,k}[n]$ and $x_{R,k}[n]$ are the left and right signals, respectively, after peripheral processing at frequency k. The *interaural cross-covariance function* is a very similar statistic in which the means are subtracted from $x_{L,k}[n]$ and $x_{R,k}[n]$ before further computation. In both cases, the ITD is typically inferred by searching for the value of m that maximizes $R[m,k]$ in each frequency channel (*e.g.,* Roman *et al.*, 2003; Brown *et al.*, 2006; Harding *et al.*, 2006; May *et al.*, 2011, 2012). Because this maximum may not occur at an integer value of m, polynomial or exponential interpolation is typically performed in the region of the maximum, with the true maximum value determined either analytically or via a grid search. In some systems (*e.g.,* Roman *et al.*, 2003; Palomäki *et al.*, 2004) the cross-correlation function is summed over frequency before the maximum is obtained. This is useful because it reduces ambiguity in identifying the true ITD of a source, particularly for larger ITDs and higher frequencies by emphasizing ITDs that are consistent over frequency as in human auditory processing (Stern *et al.*, 1988). In addition, the cross-correlation function may be “skeletonized” by replacing the normalized cross-correlation function by Gaussians located at the values of m that maximize $R[m,k]$ at each frequency (*e.g.,* Roman *et al.*, 2003; Palomäki *et al.*, 2004). This can be helpful in interpreting the responses to binaural signals that include multiple sound sources.

Systems that use STFTs as the initial stage of processing can infer ITD by calculating the phase of the product of one STFT multiplied by the complex conjugate of the other (which represents the instantaneous cross-power-spectral density function), and dividing the phase by frequency to convert to ITD (*e.g.,* Aarabi and Shi, 2004; Srinivasan *et al.*, 2006; Kim *et al.*, 2009). ITDs can also be estimated by comparing the times at which zero crossings in the signals after peripheral processing appear (Park and Stern, 2009).

In contrast, IID estimation is relatively straightforward, and is almost always estimated as the ratio of signal energies, expressed in decibels for each spectro-temporal component of the two inputs.

Mask Estimation

As noted above, the masks developed by the systems are intended to represent the extent to which a given spectro-temporal component is dominated by the target component rather than the various interfering sources or maskers in the input. The target location is either estimated by the system in initial processing (*e.g.,* Roman *et al.*, 2003; Palomäki *et al.*, 2004; May *et al.*, 2011, 2012), or by assuming a location for the target (typically directly to the front of a head or at zero ITD for two microphones). The masks are obtained by evaluating (either explicitly or implicitly) the probability of the observed ITDs and IIDs given the putative location of the sound source. For many systems these probabilities are estimated from training data, although the distributions of ITDs and especially IIDs are affected by the amount of reverberation in the environment. As noted above, the masks are either binary masks (*i.e.* equal to zero or one for each spectro-temporal component) or ratio masks (which typically take on values equal to a real number between zero and one). Because the peripheral filters are narrowband, the maxima of the interaural cross-correlation function repeat periodically along the lag axis, and the IIDs provide information that is helpful in disambiguating the cross-correlation patterns.

Another much more simple method approach is to compare the ITD estimated for each spectro-temporal component to the ITD associated with the target location, and to assign a value of one to those components that are sufficiently "close" to the the target ITD using a binary or probabilistic decision (*e.g.,* Kim *et al.*, 2009).

System Evaluation and Results

Once the mask that identifies the undistorted target components is developed, some systems use bounded marginalization (Cooke *et al.*, 2001) to recognize the target speech based on the components that are most likely to be informative (*e.g.,* Roman *et al.*, 2003; Palomäki *et al.*, 2004; May *et al.*, 2012). Other systems (*e.g.,* Kim *et al.*, 2009, 2010, 2012) reconstruct the waveform from a subset of spectro-temporal components that are deemed to be useful.

The motivations and goals of the systems considered in this subsection vary widely, making it difficult to compare them (along with other similar systems) directly. Nevertheless, a few generalizations can be made:

- Objective speech recognition and speaker identification accuracy obtained follow trends that would normally be expected: recognition accuracy degrades as SNR decreases, as the spatial separation between the target speaker and interfering sources decreases, and as the amount of reverberation increases. Recognition or identification accuracy is invariably substantially better with binaural processing compared to baseline systems that use only a single microphone.
- In situations where they can be compared directly, the use of ratio masks tends to provide greater recognition accuracy than the use of binary masks. If binary masks are used, we have found in our own work that accuracy is improved when the binary masks are smoothed over time and over frequency. The temporal smoothing can be accomplished by simply averaging

the mask values at a given frequency over a few adjacent frames. We have used "channel weighting" to accomplish the frequency smoothing, which is in essence a multiplication of the Gammatone frequency response representing each channel by the corresponding value of the binary masks and summing over frequency (*e.g.,* Kim *et al.*, 2009).

- In the single case where zero-crossing-based ITD extraction was compared to ITDs by searching for the maximum of the interaural cross-correlation function, the zero-crossing approach provided better results (Park and Stern, 2009).
- Source localization strategies in systems such as those by Roman *et al.* (2003), Palomäki *et al.* (2004), and May *et al.* (2011) appear to be effective, and their performance with multiple and moving sources should improve over time.

4.3 Robustness to Reverberation using Onset Emphasis

As noted in Sec. 2.1, many classic psychoacoustical results indicate that the auditory localization mechanism places greater emphasis on the first-arriving components of a binaural signal (*e.g.* Wallach *et al.*, 1949; Blauert, 1997; Litovsky *et al.*, 1999), a phenomenon known as the "precedence effect." More recent studies (*e.g.,* Stecker *et al.*, 2013) confirm that the lateralization of brief steady-state sounds such as tones and periodic click trains based on ITDs and IIDs appears to be strongly dominated by binaural cues contained in the initial onset portion of the sounds. In addition, Dietz *et al.* (2013) have shown that the fine-structure ITD in slow sinusoidal amplitude modulation appears to be sampled briefly during the rising-envelope phases of each modulation cycle, and is not accessed continuously over the duration of the sound.

The precedence effect is clearly valuable in maintaining a constant image location in reverberant environments when the instantaneous ITDs and IIDs produced by a sound source are likely to vary with time (Zurek *et al.*, 2004). In addition, Blauert (1983) and others have noted that the precedence effect is likely to play an important role in increasing speech intelligibility in reverberant environments. While precedence has historically been assumed to be a binaural phenomenon (*e.g.,* Lindemann (1986a)), it could also be mediated by monaural factors such as an enhancement of the onsets of envelopes of the auditory response to sound on a channel-by-channel basis at each ear (Hartung and Trahiotis, 2001).

Motivated by the potential value of onset enhancement for improved recognition accuracy in reverberation, several research groups have developed various methods of enhancing envelope onsets for improved recognition accuracy in reverberant environments. Palomäki *et al.* (2004) described an early comprehensive CASA-based binaural model that included an explicit mechanism for onset enhancement for precedence, along with other components including HRTFs, skeletonization of the cross-correlation representation, and the use of

IIDs at higher frequencies as a consistency check on the estimated binary mask. More recent algorithms that incorporate onset enhancement include the algorithm known as Suppression of Slowly-varying components and the Falling edge of the power envelope (SSF) (Kim and Stern, 2010), the temporal enhancement component of the STM algorithm (Kim *et al.*, 2011), and the SHARP algorithm (Cho *et al.*, 2016). All of these approaches incorporate nonlinear processing of the energy in the spectral envelopes to enhance transients, and they can be considered to be improved versions of the envelope enhancement approach suggested by Martin (1997) that had been used by Palomäki *et al.* (2004) and others. The temporal suppression components in Power-Normalized Cepstral Coefficients (PNCC, Kim and Stern, 2016) provide similar benefit in reverberation, but to a more limited extent.

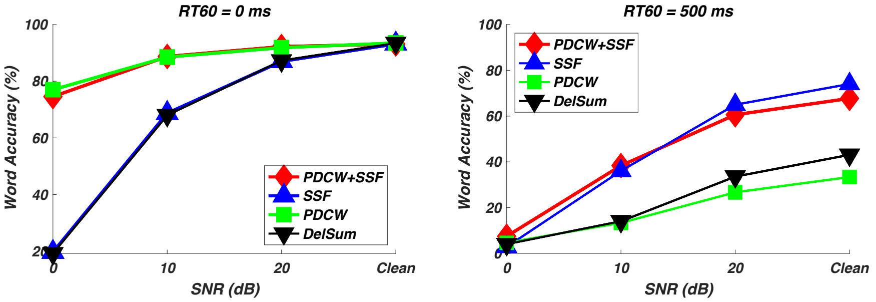

Fig. 7. Comparison of ASR accuracy obtained using the PDCW algorithm (which separates signals according to ITD) and the SSF algorithm (which enhances onsets of signal components) in the presence of additive noise and reverberation, plotted as a function of SNR. See text for further details.

Figure 7 compares selected sample recognition accuracies for the DARPA Resource Managment (RM1) task using implementations at Carnegie Mellon University of two of the approaches described above. The phase-difference channel weighting algorithm (PDCW) (Kim *et al.*, 2009)) improves ASR accuracy by separating the target speech signal from the interfering speaker according to ITD, as in other algorithms discussed in Sec. 4.2. The SSF algorithm (Kim and Stern, 2010) improves ASR accuracy by enhancing the onsets and suppressing the steady-state portions of subband components of the incoming signals, as described in this section. The data in Fig. 7 consist of a target signal directly in front of a pair of microphones in the presence of an interfering speech source at an angle of 30 degrees as well as an uncorrelated broadband noise source. The figure plots recognition accuracy obtained using delay-and-sum beamforming, PDCW, SSF, and the combination of PDCW and SSF. Results are plotted as a function of SNR for simulated reverberation times of 0 (left panel) and 500 ms (right panel). We note that the PDCW and SSF algorithms provide complementary benefits in the presence of noise and reverberation: PDCW is highly effective, even in the presence of substantial noise if there is no reverberation present, but it provides no benefit when substantial reverberation is present in the acoustical environment. SSF, on the other hand, provides substantial benefit in the presence of reverberation for the reverberation depicted, but it is ineffective in the presence of substantial additive noise. Remediation for the effects of noise is more effective than for reverberation, at least for these two algorithms.

These results suggest that the choice of which robustness approach is best in a given situation will depend on the spatial separation of target and masker components as well as the degree of reverberation in a given acoustical environment. The combination of SSF and PDCW almost always provides better performance than is observed with either algorithm by itself. While we used data from our own group for convenience in these comparisons, we believe that the use of information from ITDs (and more generally IIDs as well) to provide robustness against spatially-separated interfering sources and the use of onset enhancement to provide robustness against reverberation are generally effective across a wide range of conditions.

Pilot results from our laboratory indicate that better recognition accuracy is obtained when precedence-based onset emphasis is imposed on the input signals monaurally before binaural interaction, rather than after the binaural interaction.

4.4 Robustness to Reverberation Based on Interaural Coherence

A number of researchers have developed methods to enhance a target signal by giving greater weight to spectro-temporal components that are more "coherent" from microphone to microphone. The original motivation for much of this work was the seminal paper by Allen *et al.* (1977) who proposed that the effects of reverberation can be removed from a signal by performing a subband analysis, compensating for the ITDs observed in each frequency band, and applying a weighting in each frequency channel that is proportional to the normalized cross-correlation observed in each frequency band.

In subsequent work, Faller and Merimaa (2004) proposed that the salience of a spectro-temporal component representing a particular ITD and frequency can be characterized by a running normalized interaural cross-correlation function similar to the equation in Sec. 4.2 but updated using a moving exponential window in running time. The value of this statistic at the lag that produces the maximum interaural cross-correlation can be taken as a measure of the interaural coherence as a function of frequency.

In recent years a number of researchers have developed various models that predict the *coherent-to-diffuse energy ratio* (CDR) or the closely-related *direct-to-reverberant energy ratio* (DRR) in a given environment (*e.g.,* Jeub *et al.*, 2009, 2010, 2011a; Thiergart *et al.*, 2012; Westermann *et al.*, 2013; Zheng *et al.*, 2015). In general, the various authors use a measure similar to that proposed by Faller and Merimaa to estimate the coherent energy of the target speech and a model of the room acoustics to estimate the energy in the reverberant field. The papers differ in the assumptions that they make about the acoustics of the room, and about the geometry of the head. As a representative example we summarize the two-stage processing proposed by Jeub and colleagues (Jeub *et al.*, 2010, 2011b) for reducing the impact of reverberation. In the first stage, steering delays are imposed in the input at each frequency to compensate for differences in the path lengths from the desired source to the various microphones, and spectral subtraction is performed to suppress the effects of late reverberation. In the second stage, the residual reverberation is attenuated by a dual-channel Wiener filter derived from the coherence of the reverberant field, considering the effects of head shadowing, with the objective being suppression of the spectro-temporal components for which there is little correlation.

The systems described in the studies cited above had all been evaluated in terms of subjective or objective measures of speech quality rather than speech recognition accuracy.

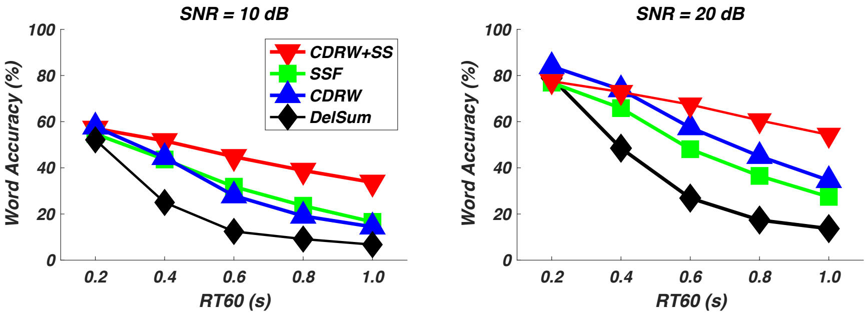

Fig. 8. Comparison of ASR accuracy obtained using the CDRW algorithm (which emphasizes signal components based on their interaural coherence) and the SSF algorithm (which enhances onsets of signal components) in the presence of additive noise and reverberation, plotted as a function of SNR. Signal-to-noise ratios are 10 dB (left panel) and 20 dB (right panel). See text for further details.

We recently completed a series of experiments (Menon, 2018) in which we compared the speech recognition accuracy obtained using a local implementation of the second stage of the algorithm of Jeub *et al.* (2011b), which enhances binaural signals based on their CDR, to the performance of the SSF algorithm described in Sec. 4.3. We refer to our implementation of spectro-temporal weighting based on CDR as CDRW (for Coherent-to-Diffuse-Ratio Weighting). Some of these results are summarized in Fig. 8, which uses similar signal sets and processing as in the data depicted in Fig. 7, except that the acoustical models used to train the ASR system were obtained using deep neural networks (DNNs). We note that both the CDRW and SSF algorithms are individually effective in reducing error rate in reverberant environments in the presence of an interfering speaker, and that SSF becomes more useful as the reverberation time increases. Moreover, the impacts of the two approaches are complementary in that best results are obtained when the two algorithms are used in combination.

4.5 EC-based Processing

The Equalization-Cancellation (EC) model of Durlach and colleagues (*e.g.* Durlach, 1963, 1972) was summarized briefly in Sec. 2.2. In developing predictions for binaural masking experiments, the EC model typically assumes that the auditory system attempts to “equalize” the masker components to the two ears by inserting ITDs and IIDs that compensate for the corresponding interaural differences that are present in the signals, and then “cancel” the masker by subtracting the signals to the two ears after equalization, leaving the target more detectable. Various investigators have proposed extensions to EC processing to accommodate the rapidly-varying fluctuations in overall ITD and IID imposed by speech-like maskers and have demonstrated that this type of processing can predict speech intelligibility (*e.g.,* Beutelmann and Brand, 2006; Beutelmann *et al.*, 2010; Wan *et al.*, 2010, 2014).

The current applications of EC-based processing to improve speech recognition accuracy differ from the traditional application of the EC model to predict binaural detection thresholds in that the equalization and cancellation operations are applied to the *target signal* rather than the masking components. This is sensible both because the SNRs tend to be greater in ASR applications than in speech threshold measurements, and because in practical applications there tends to be more useful information available *a priori* about the nature of the target speech than about the nature of the background noise and interfering signals. In the first application of this approach, Roman *et al.* (2006) employed an adaptive filter to cancel the dominant correlated signals in the two microphones, which are presumed to represent the coherent target signal. A binary mask is developed by selecting those spectro-temporal components that are the most affected by the cancellation, which are presumed to be the components most dominated by the target speech. This approach provided better ASR results for reverberated speech in the presence of multiple maskers than several other types of fixed and adaptive beamformers. Brown and Palomäki (2011) described a more sophisticated system that determines the ITD that provides maximum signal cancellation, cancels the signals to the two mics using that ITD, and again uses the absolute difference between the cancelled and uncancelled signal as an indicator of the extent to which a spectro-temporal component is dominated by the target speech. A complete signal was reconstructed using an ASR system based on bounded marginalization of the features (Cooke *et al.*, 2001), although only SRT results were provided in the paper. Mi and Colburn (2016), Mi *et al.* (2017), and Cantu (2018), among others, have developed more recent systems that enhance speech intelligibility in the presence of interfering sources based on EC principles.

4.6 Processing Using More than Two Microphones

A small number of systems have been developed that use more than two microphones. In principle, the use of more than two microphones can be useful if the additional microphones are providing new information beyond what is learned by comparing the outputs from the first two microphones. In traditional delay-and-sum beamforming and other fixed approaches such as MVDR, the amount of processing gain increases as the number microphones (or other sensors) increases. In adaptive array processing, the use of additional microphones also provides the opportunity to cancel additional interfering sources.

Over the years we have looked into the problem of "multi-aural" processing as an extension to binaural processing in various ways. In an early study, Sullivan (Sullivan and Stern, 1993; Sullivan, 1996) obtained speech recognition results using an array of up to 15 microphones, and extracting features from displays that were N-dimensional extensions of the traditional cross-correlation function after peripheral auditory processing. The system used simple peripheral processing consisting of bandpass filtering and rectification, with steering delays imposed to compensate for differences in path length from the speaker to the microphones. Performance was obtained for a number of system configurations and types of inputs. In general, recognition accuracy improved as the number of microphones was increased up to about eight. Among the results described in Sullivan (1996), it was noted that the use of post-processing with an algorithm like the vector-Taylor-series algorithm (VTS) to compensate for additive noise and spectral coloration was quite helpful. Most array-processing systems today include some sort of post-filter for similar reasons.

A second "polyaural processing" algorithm (Stern *et al.*, 2007, 2008) takes a different approach to the use of multiple microphones by mutiplying the running outputs of an N-channel binaural processor after steering delays, peripheral processing and rectification. A time-domain signal was developed in a clearly non-physiological fashion by repeating this processing for the negative outputs of each frequency channel and then adding together the positive and negative results. Because of the rectification and multiplication, this process produced a great deal of nonlinear distortion in each frequency band which was mitigated to some degree by passing the output in each channel through an additional bandpass filter at each analysis frequency. While subjective results were impressive, especially in reverberation, objective improvement in recognition accuracy was small, perhaps because of an inability to overcome the effects of the nonlinear distortion.

More recently, Moghimi considered the extent to which CASA-based selective reconstruction based on ITD using the PDCW algorithm (Kim *et al.*, 2009) can be extended to more than two microphones (Moghimi and Stern, 2014; Moghimi, 2014), along with the extent to which CASA-based ITD extraction outperforms traditional linear beamforming approaches. In general, CASA-based approaches provide greater recognition accuracy than linear beamforming with only two microphones, but as the number of microphones increases beyond four, linear processing overtakes CASA for the cases considered, primarily because the ITD information provided by the additional microphones is somewhat redundant. Moghimi and Stern (2014) demonstrate that best results are obtained when optimum linear beamforming is followed by masking based on ITD extraction according to the PDCW algorithm.

5 Binaural Processing Using Deep Learning

As we noted in Sec. 3.1, systems based on deep learning techniques are rapidly superseding conventional HMM-GMM ASR systems over the last decade, in part because of the superior ability of deep learning approaches to develop more general acoustic models. Most of the major current techniques that enable ASR using DNNs are reviewed in Hinton *et al.* (2012) as well as in the more recent book edited by Watanabe *et al.* (2017), among other resources.

Similarly, there has been great interest in the use of deep neural networks (DNNs) to perform the classifications needed to develop the binary or ratio masks to enable signal separation based on CASA principles. These approaches are reviewed comprehensively by Wang and Chen (2018), which considers (among other things) the type of mask to be employed, the choice of “training target” that is optimized in the process of training the mask classifier, the input features, the structure of the DNN used for the separation, and the methods by which the signals are separated and subsequently reconstructed.

The first system to use DNNs to separate binaural signals based on interaural differences was described by Jiang *et al.* (2014) and has components that are found in a number of similar systems. The system includes HRTFs from the KEMAR manikin, and gammatone-frequency cepstral coefficient features (GFCCs, Shao and Wang, 2008), which include 64 gammatone filters whose outputs are half-wave rectified and passed through square-root compression. ITDs are estimated using both a complete representation of the normalized cross-covariance function and a single number indicating the estimated correlation lag with maximum magnitude; IID is estimated from the subband energy ratios. Monaural GFCC features were also employed in the mask classification. A mask classifier was developed for each subband, using DNNs with two hidden layers. To avoid convergence and generalization issues with MLPs, the system was pre-trained using a restricted Boltzmann machine (RBM) (Wang and Wang, 2013). The performance of this DNN-based CASA system was compared to that of the contemporary source-separation systems DUET (Rickard, 2007) and MESSL (Mandel *et al.*, 2010), as well as systems proposed by Roman *et al.* (2003) and Woodruff and Wang (2013). Compared to the other systems considered, The DNN-based CASA system of Jiang *et al.* (2014) was found to produce substantially better approximations to the ideal binary mask that would separate the sources correctly. This system also provided improved output SNR in speech enhancement tasks. The use of the full normalized cross-correlation function (as opposed to a single numerical estimate of ITD), and with the direct inclusion of monaural features into the mask-classification process, were found to be valuable contributors to best performance. The system maintained good accuracy, and generalized to test conditions that were not included in the training for a variety of types of interfering sources and reverberant environments.

Other approaches using DNNs have been suggested as well. For example, Araki *et al.* (2015) have described the use of a denoising auto-encoder (DAE), which is trained to convert a degraded representation of a speech signal into a clean version of it. The DAE is typically structured in a “bottleneck” configuration, with at least one hidden layer that is smaller in dimensionality than the input and output layers. Estimation of a ratio mask was based on information at each frequency that included IID, ITD (as estimated from phase differences from the two inputs), and an enhanced signal was reconstructed by filtering the input using the mask that was learned by the DAE. Lowest error rates for keyword recognition in the PASCAL CHiME Speech Separation Challenge were obtained when the DNN was trained using a combination of monaural information and a location-based mask, although IID information was not useful in this particular study. Fan *et al.* (2016) described a similar system that uses a DNN with RBM-based pre-training to develop a binary mask using features that represented monaural information and IID. They observed better enhanced speech intelligibility when IIDs were extracted on a subband basis, but this system did not make use of ITD information.

Two more sophisticated binaural-based systems that separate speech using DNNs were described by Yu *et al.* (2016) and by Zhang and Wang (2017). The system of Yu *et al.* estimated ITD and IID by comparing the magnitudes and phases of the STFT components from the two microphones, along with “mixing vectors” that are obtained by combining the two monaural STFT values for each spectro-temporal component. The DNNs to estimate the mask were in the form of sparse autoencoders, which were initially trained in unsupervised fashion and later stacked to estimate the probability that each component belongs to one of several possible source directions. The system of Zhang and Wang uses both spectral and spatial features, with the spectral features obtained from the output of an MVDR beamformer with a known target location. The spatial features include ITDs represented by the complete normalized cross-correlation function along with its estimated maximum and IIDs calculated energy ratios in each frequency band. The two systems provided dramatic improvements in SNR and/or speech intelligibility for speech enhancement tasks.

The representative examples above provide merely a superficial characterization of the ever-growing body of work devoted to the development of

CASA systems using DNNs that are motivated by binaural processing to improve speech recognition accuracy. It is clear that the use of DNNs to develop the masks for speech separation systems can provide sharply improved performance compared to conventional classification techniques. This is particularly valuable because determining the spectro-temporal components of a complex input that most clearly represent the target is known to be extremely difficult, even using binaural ITD and IID information. The use of DNNs to segregate and enhance the desired target also provides impressive improvements to source localization accuracy, and to speech intelligibility, both for normal-hearing and hearing-impaired listeners. Nevertheless, this area of research is still in its infancy. For example, there is not yet a clear sense of what type of DNN architecture is best suited for mask estimation, nor is there yet a clear understanding of which monaural and binaural features are the best inputs to the DNN. Furthermore, most of the systems developed have been evaluated only in terms of measures of speech intelligibility or statistics for speech enhancement such as putative improvement in SNR. So far there have been relatively few applications of these approaches to objective tasks such as speech recognition or speaker verification. Assuming that the most effective ASR or verification systems use a DNN recognizer, it is not yet clear what is the best architecture for the purpose, nor the extent to which the form of the recognizer should be modified to accommodate missing-feature input, nor the extent to which the complete mask-estimation/recognizer-system architecture could be made more efficient or more effective by merging the two systems.

6 Summary

We have described a number of methods by which the principles of binaural processing can be exploited to provide substantial improvements in automatic speech recognition accuracy, particularly when the target speech and interfering sources are spatially separated and the degree of reverberation is moderate. In general, most of these approaches implement aspects of computational auditory scene analysis, using one of four different approaches to determine the mask which identifies the spectro-temporal components that are believed to be dominated by the target signal: direct extraction of ITDs and IIDs, onset emphasis for reverberation, exploitation of the coherent-to-diffuse ratio or related statistics, and exploitation of principles based on the E-C model. This is a particularly exciting time to be working in the application of binaural technology to automatic speech recognition because our rapidly-advancing understanding of how to develop classification techniques based on the principles of deep learning is likely to enable the realization of systems that serve their users increasingly effectively in cluttered and reverberant acoustical environments.

Acknowledgements

Preparation of this manuscript was partially supported by grants from Honeywell, Google, and Afeka University. Anjali Menon has been supported by the Prabhu and Poonam Goel Graduate Fellowship Fund and the Jack and Mildred Bowers Scholarship in Engineering. Richard Stern is deeply grateful to the many mentors, colleagues, and friends in the binaural hearing and speech recognition communities that have informed this analysis, including especially Steve Colburn, Tino Trahiotis, Bhiksha Raj, and Rita Singh. We also thank Evandro Gouvêa, Chanwoo Kim, Amir Moghimi, Hyung-Min Park, and Tom Sullivan for their many experimental contributions and general insight into these phenomena. The authors thank two anonymous reviewers for valuable comments and suggestions.

References

Aarabi, P., and Shi, G. (**2004**). “Phase-based dual-microphone robust speech enhancment,” IEEE Trans. Systems, Man, and Cybernetics, Part B **34**, 1763–1773.

Allen, J. B., Berkley, D. A., and Blauert, J. (**1977**). “Multimicrophone signal-processing technique to remove room reverberation from speech signals,” Journal of the Acoustical Society of America **62**(4), 912–915.

Allen, J. B., and Rabiner, L. R. (**1977**). “A unified approach to short-time fourier analysis and synthesis,” Proc. IEEE **65**(11), 1558–1564.

Araki, S., Hayashi, T., Delcroix, M., Fujimoto, M., Takeda, K., and Nakatani, T. (**2015**). “Exploring multi-channel features for denoissing-autoencoder-based speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 116–120.

Beutelmann, R., and Brand, T. (**2006**). “Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners,” J. Acoust. Soc. Amer. **120**, 331–342.

Beutelmann, R., Brand, T., and Kollmeier, B. (**2010**). “Revision, extension, and evaluation of a binaural speech intelligibility model,” J. Acoust. Soc. Amer. **127**, 2479–2497.

Blauert, J. (**1980**). “Modeling of interaural time and intensity difference discrimination,” in *Psychophysical, Physiological, and Behavioural Studies in Hearing*, edited by G. van den Brink and F. Bilsen (Delft University Press, Delft), pp. 412–424.

Blauert, J. (**1983**). “Review paper: Psychoacoustic binaural phenomena,” in *Hearing – Physiologica Bases and Psychophysics*, edited by R. Klinke and R. Hartmann (Springer-Verlag, Heidelberg), pp. 182–189.

Blauert, J. (**1997**). *Spatial hearing: the Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA), revised edition.

Blauert, J., and Cobben, W. (**1978**). “Some considerations of binaural cross-correlation analysis,” Acustica **39**, 96–103.

Bodden, M. (**1993**). “Modelling human sound-source localization and the cocktail party effect,” Acta Acustica **1**, 43–55.

Bodden, M., and Anderson, T. R. (**1995**). “A binaural selectivity model for speech recognition,” in *Proceedings of Eurospeech 1995* (European Speech Communication Association).

Boll, S. F. (**1979**). “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech and Signal Processing **27**(2), 113–120.

Bourlard, H., and Morgan, N. (**1994**). *Connectionist Speech Recognition: A hybrid approach* (Kluwer Academicc Publishers).

Braasch, J. (**2005**). “Modelling of binaural hearing,” in *Communication Acoustics*, edited by J. Blauert (Springer-Verlag, Berlin), Chap. 4, pp. 75–108.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**a). “Binaural processing model based on contralateral inhibition. I. Model structure,” Journal of the Acoustical Society of America **110**, 1074–1088.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**b). “Binaural processing model based on contralateral inhibition. II. Dependence on spectral parameters,” Journal of the Acoustical Society of America **110**, 1089–1103.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**c). “Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters,” Journal of the Acoustical Society of America **110**, 1125–1117.

Bregman, A. S. (**1990**). *Auditory scene analysis* (MIT Press, Cambridge, MA).

Brown, G. J., and Cooke, M. P. (**1994**). “Computational auditory scene analysis,” Computer Speech and Language **8**, 297–336.

Brown, G. J., Harding, S., and Barker, J. P. (**2006**). “Speech separation based on the statistics of binaural auditory features,” in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. V, pp. 949 – 952.

Brown, G. J., and Palomäki, K. J. (**2011**). “A computational model of binaural speech recognition: Ro;e of across-frequency vs. within-frequency processing and internal noise,” Speech Communication **53**, 924–940.

Burkhard, M. D., and Sachs, R. M. (**1975**). “Anthroponetric manikin for acoustic research,” Journal of the Acoustical Society of America **58**, 214–222.

Cantu, M. (**2018**). “Sound source segregation of multiple concurrent talkers via short-time target cancellation,” Ph.D. thesis, Boston University.

Cho, B. J., Kwon, H., Cho, J.-W., Kim, C., Stern, R. M., and Park, H.-M. (**2016**). “A subband-based stationary-component suppression method using harmonics and power ratio for reverberant speech recognition,” IEEE Signal Processing Letters **23**(6), 780–784.

Colburn, H. S. (**1969**). “Some physiological limitations on binaural performance,” Ph.D. thesis, Massachusetts Institute of Technology.

Colburn, H. S. (**1973**). “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 **54**, 1458–1470.

Colburn, H. S., and Durlach, N. I. (**1978**). “Models of binaural interaction,” in *Hearing*, edited by E. C. Carterette and M. P. Friedmann, **IV** of *Handbook of Perception* (Academic Press, New York), Chap. 11, pp. 467–518.

Colburn, H. S., and Kulkarni, A. (**2005**). “Models of sound localization,” in *Sound Source Localization*, edited by R. Fay and T. Popper, Springer Handbook of Auditory Research (Springer-Verlag), Chap. 8, pp. 272–316.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**). “Robust automatic speech recognition with missing and unreliable acoustic data,” Speech Communication **34**, 267–285.

Cooke, M. P., and Ellis, D. P. W. (**2001**). “The auditory organization of speech and other sources in listeners and computational models,” Speech Communication **35**, 141–177.

Davis, S. B., and Mermelstein, P. (**1980**). “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech and Signal Processing **28**, 357–366.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society, Series B **39**, 1–38.

DeSimio, M. P., Anderson, T. R., and Westerkamp, J. J. (**1996**). “Phoneme recognition with a model of binaural hearing,” IEEE Trans. on Speech and Audio Processing **4**, 157–166.

Dietz, M., Lestang, J. H., Majdak, P., Stern, R. M., Marquardt, T., Ewert, S. D., Hartmann, W. M., and Goodman, D. F. M. (**2017**). “A framework for testing and comparing binaural models,” Hearing Research **360**, 92–106.

Dietz, M., Marquardt, T., Salminen, N. H., and McAlpine, D. (**2013**). “Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds,” Proc. Natl. Acad. Sci. U.S.A. **110**, 15151—15156.

Domnitz, R. H., and Colburn, H. S. (**1976**). “Analysis of binaural detection models for dependence on interaural target parameters,” J. Acoust. Soc. Amer. **59**, 599–601.

Domnitz, R. H., and Colburn, H. S. (**1977**). “Lateral position and interaural discrimination,” Journal of the Acoustical Society of America **61**, 1586–1598.

Droppo, J. (**2013**). “Feature compensation,” in *Techniques for Noise Robustness in Automatic Speech Recognition*, edited by T. Virtanen, B. Raj, and R. Singh (Wiley), Chap. 9.

Durlach, N. I. (**1963**). “Equalization and cancellation theory of binaural masking level differences,” Journal of the Acoustical Society of America **35**(8), 1206–1218.

Durlach, N. I. (**1972**). “Binaural signal detection: Equalization and cancellation theory,” in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias, **2** (Academic Press, New York), pp. 369–462.

Durlach, N. I., and Colburn, H. S. (**1978**). “Binaural phenomena,” in *Hearing*, edited by E. C. Carterette and M. P. Friedman, **IV** of *Handbook of Perception* (Academic Press, New York), Chap. 10, pp. 365–466.

Faller, C., and Merimaa, J. (**2004**). “Sound localization in complex listening situations: Selection of binaural cues based on interaural coherence,” Journal of the Acoustical Society of America **116**(5), 3075–3089.

Fan, N., Du, J., and Dai, L.-R. (**2016**). “A regression approach to binaural speech segregation via deep neural networks,” in *Proc. IEEE International Symposium on Chinese Spoken Language Processing*, pp. 116–120.

Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G. W. (**1985**). “Computer-steered microphone arrays for sound transduction in large rooms,” J. Acoustic. Soc. Amer. **78**, 1508–1518.

Gaik, W. (**1993**). “Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling,” Journal of the Acoustical Society of America **94**, 98–110.

Gardner, B., and Martin, K. (**1994**). “HRTF measurements of a KEMAR dummy-head microphone,” Technical Report 280 , available online at http://sound.media.mit.edu/KEMAR.html.

Gilkey, R. H., and Anderson, T. A., eds. (**1997**). *Binaural and Spatial Hearing in Real and Virtual Environments* (Psychology Press).

Gold, B., Morgan, N., and Ellis, D. (**2011**). *Speech and Audio Signal Processing*, 2 ed. (Wiley Interscience).

Goodfellow, I., Bengio, Y., and Courville, A. (**2016**). *Deep Learning* (MIT Press).

Harding, S., Barker, J., and Brown, G. J. (**2006**). "Mask estimation for missing data speech recognition based on statistics of binaural interaction," IEEE Trans. on Speech and Audio Processing **14**, 58–67.

Hartung, K., and Trahiotis, C. (**2001**). "Peripheral auditory processing and investigations of the "precedence effect" which utilize successive transient stimuli," J. Acoust. Soc. Amer. **110**(3), 1505–1513.

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

Haykin, S. (**2018**). *Neural Networks And Learning Machines*, third ed. (Springer).

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) anlysis of speech," J. Acoustic. Soc. Amer. **87**(4), 1738–1752.

Hermansky, H., Ellis, D. P. W., and Sharma, S. (**2000**). "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. IEEE ICASSP*, pp. 1635–1638.

Hermansky, H., and Morgan, N. (**1994**). "RASTA processing of speech," IEEE Transactions on Speech and Audio Processing **2**, 578–589.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., MOhamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (**2012**). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine **29**, 82–97.

Jeffress, L. A. (**1948**). "A place theory of sound localization," J. Comp. Physiol. Psych. **41**, 35–39.

Jeub, M., Dorbecker, M., and Vary, P. (**2011**a). "Semi-analytical model for the binaural coherence of noise fields," IEEE Signal Processing Letters **18**(3), 197–200.

Jeub, M., Nelke, C., Beaugeant, C., and Vary, P. (**2011**b). "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Proc.* 19^{th} *European Signal Processing Conference*.

Jeub, M., Schafer, M., Esch, T., and Vary, P. (**2010**). "Model-based dereverberation preserving binaural cues," IEEE Transactions on Audio, Speech, and Language Processing **18**(7), 1732–1745.

Jeub, M., Schafer, M., and Vary, P. (**2009**). "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc.* 16^{th} *International Conference on Digital Signal Processing*, pp. 1–5.

Jiang, Y., Wang, D., Liu, R., and Feng, Z. (**2014**). "Binaural classification for reverberant speech segregation using deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**(12), 2112–2121.

Johnson, D. H., and Dudgeon, D. E. (**1993**). *Array Signal Processing: Concepts and Techniques* (Prentice-Hall, Englewood Cliffs NJ).

Kates, J. M. (**1991**). "A time-domain digital cochlear model," IEEE Trans. on Signal Processing **39**, 2573–2592.

Kim, C., Khawand, C., and Stern, R. M. (**2012**). "Two-microphone source separation algorithm based on statistical modeling of angle distributions," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Kyoto, Japan.

Kim, C., Kumar, K., Raj, B., and Stern, R. M. (**2009**). "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain," in *Proc. Interspeech*.

Kim, C., Kumar, K., and Stern, R. M. (**2011**). "Binaural sound source separation motivated by auditory processing," in *Proc. Interspeech*, Prague, Czech Republic, Vol. 23, pp. 780–784.

Kim, C., and Stern, R. M. (**2010**). "Nonlinear enhancement of onset for robust speech recognition," in *Proc. Interspeech*, Makuhari, Japan.

Kim, C., and Stern, R. M. (**2016**). "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," IEEE Trans. on Audio, Speech, and Language Proc. **24**(7), 1315–1329.

Kim, C., Stern, R. M., Eom, K., and Kee, J. (**2010**). "Automatic selection of thresholds for signal separation algorithms based on interaural delay," in *Proc. Interspeech*, Makuhari, Japan.

Kohonen, T. (**1989**). "The neural phonetic typewriter," IEEE Computer Magazine 11–22.

Kolrausch, A., Braasch, J., Kolossa, D., and Blauert, J. (**2013**). "An introduction to binaural processing," in *The Technology of Binarual Listening*, edited by J. Blauert, Modern Acoustics and Signal Processing (Springer, Berlin).

Kumatani, K., McDonough, J., and Raj, B. (**2012**). "Microphone array processing for robust speech recognition," IEEE Signal Processing Magazine **29**(6), 127–140.

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

Lindemann, W. (**1986**b). "Extension of a binaural cross-correlation model by contralateral inhibition. II. the law of the first wavefront," Journal of the Acoustical Society of America **80**, 1623–1630.

Lippmann, R. P. (**1987**). "An introduction to computing with neural nets," IEEE ASSP Magazine **4**(2), 4–22.

Lippmann, R. P. (**1989**). "Review of neural networks for speech recognition," Neural Computation **1**(1), 1–38.

Litovsky, R. Y., Colburn, S. H., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," Journal of the Acoustical Society of America **106**, 1633–1654.

Lyon, R. F. (**1984**). "Computational models of neural auditory processing," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing of the International Conference on Acoustics, Speech and Signal Processing*, pp. 36.1.1–36.1.4.

Mandel, M. I., Weiss, R. J., and Ellis, D. P. W. (**2010**). "Model-based expectation-maximization source separation and localization," IEEE Trans. on Audio, Speech, and Language Proc. **18**(2), 382–394.

Martin, K. D. (**1997**). "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE Mohonk Workshop on Applications of Signal Processing to Acoustics and Audio.*

May, T., Par, S. V. D., and Kohlrausch, A. (**2012**). "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," IEEE Transactions on Audio, Speech, and Language Processing **20**, 108–121.

May, T., van de Par, S., and Kohlrausch, A. (**2011**). "A probabilistic model for robust localization based on a binaural auditory front-end," IEEE Transactions on Audio, Speech, and Language Processing **19**(1), 1–13.

McGovern, S. G. (**2004**). "Room impulse response generator (MATLAB code)," Http://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator.

Mehrgardt, S., and Mellert, V. (**1977**). "Transformation charactersitics of the external human ear," Journal of the Acoustical Society of America **61**, 1567–1576.

Menon, A. (**2018**). "Robust recognition of binaural speech signals using techniques based on human auditory processing," Ph.D. thesis, Carnegie Mellon University.

Mi, J., and Colburn, H. S. (**2016**). "A binaural grouping model for predicting speech intelligibility in multitalker environments," Trends in Hearing **20**, 1–12.

Mi, J., Groll, M., and Colburn, H. S. (**2017**). "Comparison of a target-equalization-cancellation approach and a localization approach to source separation," J. Acoust. Soc. Amer **142**(5), 2933–2941.

Miao, Y., and Metze, F. (**2017**). "End-to-end architectures for speech recognition," in *New Era for Robust Speech Recognition: Exploiting Deep Learning*, edited by S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey (Springer International Publishing), pp. 299–323.

Mitra, V., Franco, H., Stern, R., Hout, J. V., Ferrer, L., Graciarena, M., Wang, W., Vergyri, D., Alwan, A., and Nansen, J. H. L. (**2017**). "Robust features in deep learning-based speech recognition," in *New Era for Robust Speech Recognition: Exploiting Deep Learning*, edited by S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey (Springer International Publishing), pp. 183–212.

Moghimi, A. (**2014**). "Array-based spectro-temporal masking for automatic speech recognition," Ph.D. thesis, Carnegie Mellon University.

Moghimi, A., and Stern, R. M. (**2014**). "Post-masking: A hybrid approach to array processing for speech recognition," in *Proc. Interspeech*.

Moore, B. C. J. (**2012**). *An introduction to the psychology of hearing, Sixth edition*, fifth ed. (Emerald Group Publishing Ltd., Bingley UK, London).

Moreno, P. J., Raj, B., and Stern, R. M. (**1996**). "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 733–736.

Nielsen, M. (**2016**). *Neural Networks and Deep Learning* (http://neuralnetworksanddeeplearning.com/).

Osman, E. (**1971**). "A correlation model of binaural masking level differences," J. Acoust. Soc. Amer. **50**, 1494–1511.

Palomäki, K. J., Brown, G. J., and Wang, D. L. (**2004**). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," Speech Communication **43**(4), 361–378.

Park, H.-M., and Stern, R. M. (**2009**). "Spatial separation of speech signals using continuously-variable weighting factors estimated from comparisons of zero crossings," Speech Communication Journal **51**(1), 15–25.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). "An efficient auditory filterbank based on the gammatone function,," Applied Psychology Unit (APU) Report 2341, Cambridge, UK.

Rabiner, L. R. (**1989**). "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE **77**(2), 257–286.

Rabiner, L. R., and Juang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall).

Raj, B., Seltzer, M. L., and Stern, R. M. (**2004**). "Reconstruction of missing features for robust speech recognition," Speech Communication **43**(4), 275–296.

Raj, B., and Stern, R. M. (**2005**). "Missing-feature approaches in speech recognition," IEEE Signal Processing Magazine **22**(5), 101–115.

Rickard, S. (**2007**). "The duet blind source separation algorithm," in *Blind Speech Separation*, edited by S. Makino, T. Lee, and H. E. Sawada (Springer-Verlag (New York)).

Roman, N., Srinivasan, S., and Wang, D. (**2006**). "Binaural segregation in multisource reverberant environments," J. Acoust. Soc. Amer. **120**(4040–4051).

Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," Journal of the Acoustical Society of America **114**(4), 2236–2252.

Rosenblatt, R. (**1959**). *Principles of Neurodynamics* (Spartan Books, New York).

Schroeder, M. R. (**1977**). "New viewpoints in binaural interactions," in *Psychophysics and Physiology of Hearing*, edited by E. F. Evans and J. P. Wilson (Academic Press (London)), pp. 455–467.

Shamma, S. A., Shen, N., and Gopalaswamy, P. (**1989**). "Binaural processing without neural delays," Journal of the Acoustical Society of America **86**, 987–1006.

Shao, Y., and Wang, D. L. (**2008**). "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 1589–1592.

Srinivasan, S., Roman, M., and Wang, D. (**2006**). "Binary and ratio time-frequency masks for robust speech recognition," Speech Comm. **48**, 1486–1501.

Stecker, G. C., Ostreicher, J. D., and Brown, A. D. (**2013**). "Temporal weighting functions for interaural time and level differences. iii. temporal weighting for lateral position judgments," J. Acoust. Soc. Amer **134**, 1242–1252.

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

Stern, R. M., Gouvêa, E., and Thattai, G. (**2007**). "'Polyaural' processing for automatic speech recognition in degraded environments," in *Proc. Interspeech.*

Stern, R. M., Gouvêa, E. B., Kim, C., Kumar, K., and Park, H.-M. (**2008**). "Binaural and multiple-microphone signal processing motivated by auditory perception," in *Proceedings of the HSCMA Joint Workshop on Hands-free Speech Communication and Microphone Arrays.*

Stern, R. M., and Trahiotis, C. (**1995**). "Models of binaural interaction," in *Hearing*, edited by B. C. J. Moore, Handbook of Perception and Cognition, 2 ed. (Academic (New York)), Chap. 10, pp. 347–386.

Stern, R. M., and Trahiotis, C. (**1996**). "Models of binaural perception," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. Gilkey and T. R. Anderson (Lawrence Erlbaum Associates), Chap. 24, pp. 499–531.

Stern, R. M., Wang, D., and Brown, G. J. (**2006**). "Binaural sound localization," in *Computational Auditory Scene Analysis*, edited by D. Wang and G. J. Brown (Wiley-IEEE Press), Chap. 5.

Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (**1988**). "Lateralization of complex binaural stimuli: a weighted image model," Journal of the Acoustical Society of America **84**, 156–165.

Stevens, S. S., Volkman, J., and Newman, E. (**1937**). "A scale for the measurement of the psychological magnitude pitch," J. Acoustic. Soc. Amer. **8**(3), 185–190.

Stockham, T. G., Cannon, T. M., and Ingrebretsen, R. B. (**1975**). "Blind deconvolution through digital signal processing," Proc. IEEE **63**(4), 678–692.

Sullivan, T. M. (**1996**). "Multi-microphone correlation-based processing for robust automatic speech recognition," Ph.D. thesis, Carnegie Mellon University.

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

Thiergart, O., Del Galdo, G., and Habets, E. A. (**2012**). "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, IEEE, pp. 309–312.

Trahiotis, C., Bernstein, L. R., Stern, R. M., and Buell, T. N. (**2005**). "Interaural correlation as the basis of a working model of binaural processing: An introduction," in *Sound Source Localization*, edited by R. Fay and T. Popper, Springer Handbook of Auditory Research (Springer-Verlag, Heidelberg), Chap. 7, pp. 238–271.

Van Trees, H. L. (**2004**). *Detection, Estimation, and Modulation Theory: Optimum Array Processing* (John Wiley & Sons).

Virtanen, T., Raj, B., and Singh, R., eds. (**2012**). *Noise-Robust Techniques for Automatic Speech Recognition* (Wiley).

Wallach, H. W., Newman, E. B., and Rosenzweig, M. R. (**1949**). "The precedence effect in sound localization," American Journal of Psychology **62**, 315–337.

Wan, R., Durlach, N. I., and Colburn, H. S. (**2010**). "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," J. Acoust. Soc. Amer. **128**, 3678–3690.

Wan, R., Durlach, N. I., and Colburn, H. S. (**2014**). "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers," J. Acoust. Soc. Amer. **136**, 768–776.

Wang, D., and Brown, G. J., eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE Press).

Wang, D. L., and Chen, J. (**2018**). "Supervised speech separation based on deep learning: An overview," IEEE Transactions on Audio, Speech, and Language Proc. **26**, 1702–1726.

Wang, Y., and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Transactions on Audio, Speech, and Language Proc. **21**, 1381–1390.

Watanabe, S., Delcroix, M., Metze, F., and Hershey, J. R., eds. (**2017**). *New Era for Robust Speech Recognition: Exploiting Deep Learning* (Springer International).

Westermann, A., Buchholz, J. M., and Dau, T. (**2013**). "Binaural dereverberation based on interaural coherence histograms a," The Journal of the Acoustical Society of America **133**(5), 2767–2777.

Wightman, F. L., and Kistler, D. J. (**1989**a). "Headphone simulation of free-field listening. I: Stimulus synthesis," J. Acoustic. Soc. Amer. **85**, 858–867.

Wightman, F. L., and Kistler, D. J. (**1989**b). “Headphone simulation of free-field listening. II: Psychophysical validation,” Journal of the Acoustical Society of America **87**, 868–878.

Wightman, F. L., and Kistler, D. J. (**1999**). “Resolution of front–back ambiguity in spatial hearing by listener and source movement,” J. Acoust. Soc. Amer. **105**(5), 2841–2853.

Woodruff, J., and Wang, D. L. (**2013**). “Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues,” IEEE Trans. on Audio, Speech, and Language Proc. **21**, 806–815.

Yost, W. A. (**1981**). “Lateral position of sinusoids presented with intensitive and temporal differences,” Journal of the Acoustical Society of America **70**, 397–409.

Yost, W. A. (**2013**). *Fundamentals of Hearing: An Introduction, Fifth Edition*, 5 ed. (Academic Press, Burlington MA).

Yu, Y., Wang, W., and Han, P. (**2016**). “Localization based stero speech source separation using probabilistic time-frequency masking and deep neural networks,” EURASIP Journal on Audio, Speech, and Music Processing **2016**, 1–18.

Zhang, X., Heinz, M. G., Bruce, I. C., and Carney, L. H. (**2001**). “A phenomenological model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppresion,” Journal of the Acoustical Society of America **109**, 648–670.

Zhang, X., and Wang, D. (**2017**). “Deep learning based binaural speech separation in reverberant environments,” IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(5), 1075–1084.

Zheng, C., Schwarz, A., Kellermann, W., and Li, X. (**2015**). “Binaural coherent-to-diffuse-ratio estimation for dereverberation using an ITD model,” in *Proc.* 23^{rd} *European Signal Processing Conference (EUSIPCO)*, pp. 1048–1052.

Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and Carney, L. H. (**2009**). “A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics,” J. Acoust. Soc. Amer **125**, 2390–2412.

Zurek, P. M. (**1993**). “Binaural advantages and directional effects in speech intelligibility,” in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, Boston).

Zurek, P. M., Freyman, R. L., and Balakrishnan, U. (**2004**). “Auditory target detection in reverberation,” Journal of the Acoustical Society of America **115**(4), 1609–1620.