Speech segregation based on sound localization

Nicoleta Roman, DeLiang Wang
Department of Computer and Information Science and Center for Cognitive Science
The Ohio State University, Columbus, OH 43210, USA
Email: {niki, dwang}@cis.ohio-state.edu

Guy J. Brown
Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
Email: g.brown@dcs.shef.ac.uk

Abstract

We study the cocktail-party effect, which refers to the ability of a listener to attend to a single talker in the presence of adverse acoustical conditions. It has been observed that this ability improves in the presence of binaural cues. In this paper, we explore a technique for speech segregation based on sound localization cues. The auditory masking phenomenon motivates an "ideal" binary mask in which time-frequency regions that correspond to the weak signal are canceled. In our model we estimate this binary mask by observing that systematic changes of the interaural time differences and intensity differences occur as the energy ratio of the original signals is modified. The performance of our model is comparable with results obtained using the ideal binary mask and it shows a large improvement over existing pitch-based algorithms.

1 Introduction

The field of Computational Auditory Scene Analysis (CASA) is preoccupied with solving the sound source separation problem, with emphasis on modeling auditory scene analysis in humans. Sound sources may differ in location, fundamental frequency, or the patterns of envelope modulation in different frequency bands. These represent potential grouping cues used in a bottom-up process (so-called primitive process) in order to organize components with a common origin into a single stream. The primary grouping cue used in most CASA systems is fundamental frequency (F0) - this works well only for parts of the speech signal that contain voiced components. On the other hand, binaural cues have the advantage of being independent of the signal structure and can be used for sequential integration across both voiced and unvoiced components.

The main cues used by the binaural auditory system are interaural time differences and interaural intensity differences (Lord Rayleigh [16]). Psychoacoustic experiments show that interaural time differences (ITD) are most effective at low frequencies (<1.5kHz) and interaural intensity differences (IID) dominate the high frequency range. Jeffress described a simple and intuitive mechanism that performs a running interaural cross-correlation by means of a neurophysiologically plausible network [13]. This mechanism accounts for the lateral displacement of the auditory event from the median plane when an ITD is present. Mechanisms additional to the cross-correlation model have been proposed to simulate auditory event localization based on both on ITD and IID [4], [5], [8].

Models of binaural hearing have already been used for sound source separation [9], [6]. The main underlying observation in most of the existing models is that the auditory event corresponding to a desired sound source undergoes systematic changes due to the interfering noise. Our model attempts to quantify those changes by collecting statistics using a corpus of mixtures of speech and interfering noise.

In the next section we briefly describe the architecture of our model. In the third section we introduce our method for estimating the ideal binary mask – a binary matrix of time and frequency. Our goal is to estimate this ideal mask based on measured ITDs and IIDs across frequency bands. The fourth section presents our results and observations by comparing our model with an existing pitch-based algorithm.

2 Model Architecture

The input to our model consists of two signals: one speech signal and one interfering noise with sampling frequency of 44.1 kHz. The model consists of five stages that are presented in the following subsections and Section 3, and shown schematically in Fig. 1.

2.1 Eardrum signals

For a free-field presentation, the acoustic signals at the eardrums consist entirely of direct sound from the sound source (no echoes or reverberations are assumed in the current model). Binaural signals for such free-field sources can be obtained by convolving input signals with

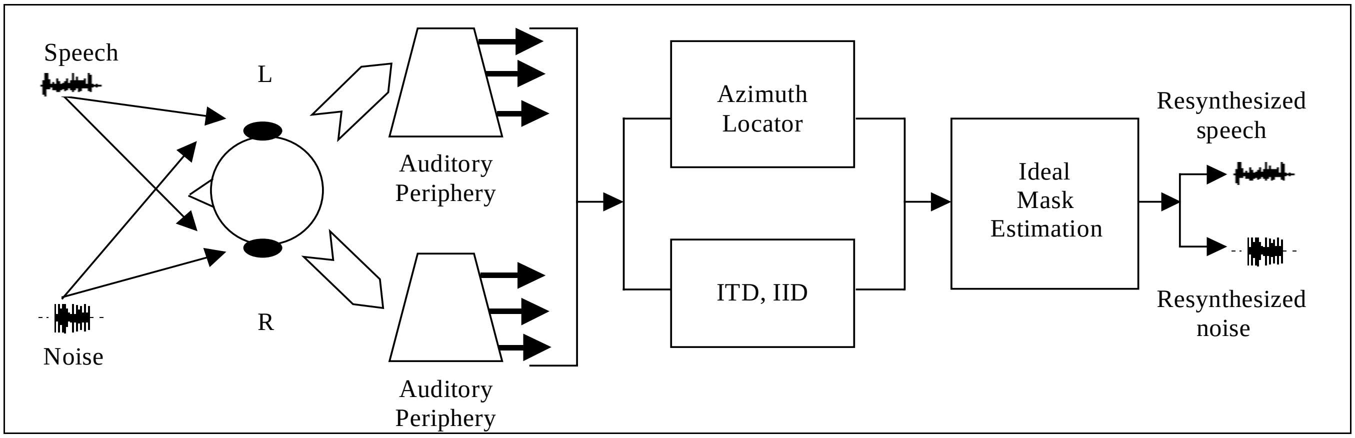

Figure 1: Schematic diagram of the model. The model processes input from two sound sources with different locations (different azimuths). First stage: binaural signals are obtained by convolving the input signals with HRIRs. Second stage corresponds to the auditory periphery simulation: cochlear filtering, half-wave rectification to simulate auditory nerve firing and square root to simulate saturation effects. Third stage: azimuth localization of the two sound sources and computation of IIDs and ITDs across frequency bands. Fourth stage: estimation of the ideal binary mask. Fifth stage: the resynthesis path allows reconstruction of the separated signals [10].

measured head related impulse responses (HRIR) from a KEMAR dummy head, which gives realistic filtering due to the external ear [7]. Two sound sources are simulated: one corresponds to speech and another to interfering noise. The corresponding left and right signals for the two sources are summed at the eardrums.

Location dependent ITDs and IIDs arise naturally in a free field environment due to diffraction, scattering, interference and resonance effects. The range of ITDs is reported to be up to 800 μs . For IIDs, as much as 30 dB level differences can be obtained for high frequencies [2].

2.2 Auditory periphery

Peripheral auditory processing is simulated using a bank of 128 gammatone filters as described in [10], [3]. In addition, the gains of the gammatone filters are adjusted in order to simulate the middle ear transfer function [15]. In the final stage of the peripheral model, the output of each gammatone filter is half-wave rectified in order to simulate the firing probabilities of nerve fibers. . Saturation effects are modeled by taking the square root of the signal.

2.3 Azimuth Locator

Current models of azimuth localization use, as a starting point, Jeffress's cross-correlation mechanism [13] ([4], [5], [6], [8], [9]). Cross-correlation provides excellent time delay estimation for broadband signals and narrowband stimuli in the low frequency range. However, for periodic waveforms it can present ambiguous peaks at intervals of the fundamental frequency. In our model the cross-correlation is implemented by computing cross-correlation coefficients at time delays equally distributed in the plausible range from –1 ms to 1 ms for all frequency channels. ITDs across frequency bands are estimated at the position $\tau^{i}_{\max}$ of the absolute maximum of the cross-correlation function in the *i*th channel. In a training phase, we derive frequency-dependent nonlinear transformations to map the time-delay axis onto an azimuth axis. Diffraction effects introduce weak frequency dependences for ITDs (Fig. 2A). The functions are monotonic, being sigmoidal at low frequencies (where diffraction effects are greater) and increasingly linear at high frequencies.

Cross-correlation provides inconsistent results when two acoustical sources are present (Fig. 2B). For frequency channels that are dominated by one source, activity is observed near the true location. For frequency–time regions where the two sources overlap the peak deviates, generally being closer to the louder source. Peaks at both locations can occur in high frequency channels – this ambiguity is due to the periodicity of the cross-correlation function. Hence, under certain conditions (sufficient channels where no overlapping occurs) an estimate of the two sound source locations can be obtained at every time frame (Fig. 2C). Since we assume fixed locations in time, a summary across time and frequency weighted by the energy will produce two peaks corresponding to the two true locations. Further stages of our model assume perfect localization.

3. Binary Mask Estimation

3.1 Ideal Binary Mask

The auditory masking effect states that for narrowband stimuli with close frequencies (same critical band) the stronger signal masks the weaker one [14]. In light of this phenomenon we estimate an ideal binary mask by

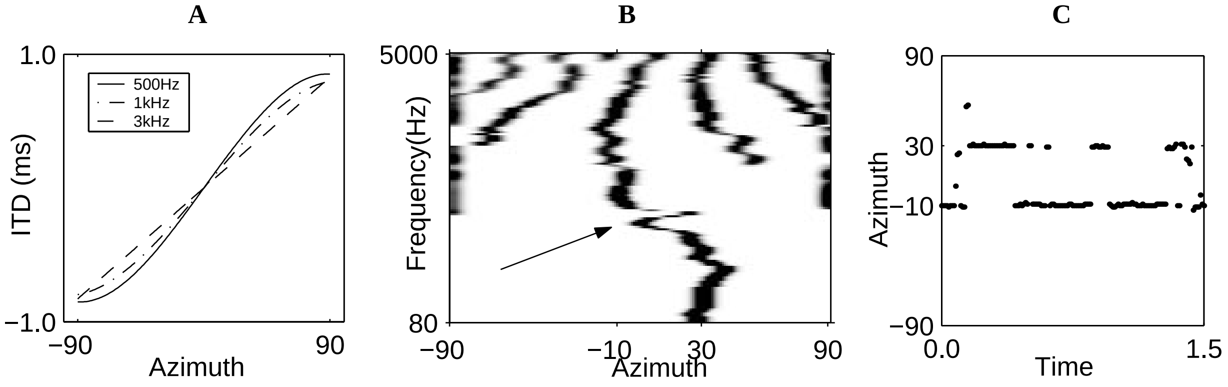

Figure 2. A: Functions relating azimuth to ITD for three channels of the auditory model with CFs of 500Hz, 1kHz, 3kHz. **B:** Cross-correlation for a mixture of male speech at 30° degrees and female speech at -10° degrees (128 channels, time frame 40 (0.4 ms)). **C:** Azimuth localization for the two most predominant sources in a mixture of speech at 30° and telephone ringing at -10° updated at every 10 ms.

comparing the energies of the original signals that arrive at the better ear (closer to the speech source). The idea is to pass time-frequency regions where speech is predominant and mask the other regions. Recent investigations also show that robust results can be obtained using similar binary masks as front-end processors to automatic speech recognizers [12].

We compute energy ratios using the following formula:

$$E_i = \frac{\sum_t s_i^2(t)}{\sum_t s_i^2(t) + \sum_t n_i^2(t)} \qquad (1)$$

where s_i and n_i refer to the output of the *i*th gammatone filter for speech and noise, respectively. The masking coefficients of the ideal binary mask are set to 1 whenever the corresponding energy ratio exceeds the threshold 0.5 (speech spectrum dominates the noise spectrum), and 0 otherwise.

Our approach is to design a method that approximates this ideal binary mask when correct information about the locations of the two sound sources is extracted from the azimuth locator.

3.2 Pure Tones

A psychophysical motivation for our model is the "summing localization" phenomenon [2]. The classical experiment to describe this phenomenon uses two loudspeakers positioned symmetrically in front of the subject. If both loudspeakers are driven with identical signals an auditory event is perceived in the median plane. By introducing a time delay or an intensity difference between the two signals, the perceived position of the auditory event moves away from the median plane toward the loudspeaker that emits earlier or is louder.

For a frequency band, local information about the location of the auditory event is extracted using ITDs for low frequencies (<1.5 kHz) and IIDs for high frequencies. By extrapolating the summing localization results to our bandfiltered signals we expect to observe a correlation between ITDs and IIDs corresponding to the *i*th channel and the energy ratios E_i.

For low frequency channels, the gammatone output is a narrowband signal for which we can neglect the IID. To start, we analyze a simple mathematical model for two sources of pure tones. The theoretical cross-correlation function for this system is given by the following formula:

$$\begin{aligned} c(\tau) = {} & \frac{A_1^2}{2}\cos(\omega(\tau - d_1)) + \frac{A_1^2}{2}\cos(\omega(\tau - d_2)) + \\ & + \frac{A_1 A_2}{2}\cos(\omega(\tau - \frac{d_1 + d_2}{2}))\cdot coeff, \qquad (2) \\ coeff = {} & \cos(\Delta\varphi + \omega\frac{d_2 - d_1}{2}) \end{aligned}$$

where A_i is the amplitude of the pure sine wave and d_i represents interaural time delay for the *i*th source; $\Delta\varphi$ depends on phase differences between the initial signals and those due to the arrival times of the signals at the left ear. As a consequence, *coeff* is assumed to be random variable in the interval [-1, 1]. We observe that a systematic change in the relative amplitude results in a systematic shift of the peak location in the cross-correlation function. By observing the deviation of the peak location $\tau_{\max}$ from the middle location between the two sources we can estimate which signal is stronger:

$$\tau_{\max} > \frac{d_1 + d_2}{2} \Leftrightarrow A_1 > A_2 \qquad (3)$$

We extrapolate these results for the low frequency channels and conclude that there exists a correlation between the energy ratios E_i and the time delays that corresponds to the maximum in the cross-correlation pattern for the *i*th channel.

3.3 Method

Energy ratios E_i, IIDs and ITDs are computed at a specific time frame using a time window of 20 ms and are updated at every 10 ms across all frequency bands. ITDs correspond to the time location of the maximum in the cross-correlation pattern (we have already seen that in low frequency channels one unambiguous peak is obtained). IIDs are computed based on the following formula:

$$L_i = 20 \log_{10} \frac{\sum_t l_i^2}{\sum_t r_i^2} \qquad (4)$$

where l_i, r_i refer to the auditory periphery output for the *i*th frequency channel.

In a learning stage, for every pair of azimuth angles (θ_1, θ_2) we collect statistics of the relationships between binaural cues and the energy ratios E_i across all critical bands. The corpus has 100 mixtures obtained from 10 speech signals located at θ_1 and 10 noise intrusions located at θ_2 [11]. The correlation between ITD and the energy ratio E_i proves to be most effective at low frequencies (<1.5kHz). For high frequency channels (>1.5kHz) we evaluate the performance of two cues: ITDs based on the envelopes of the signals in order to avoid the multiple peak problem and IIDs. IIDs are more reliable for high frequencies. Our computational observations match psychoacoustical results.

In Fig. 3 we display statistics for the relationships between binaural cues and energy ratios when $\theta_1 = 30^\circ$, $\theta_2 = -10^\circ$. We observe that the ITD and the IID values undergo relatively smooth changes with the energy ratio. The particular orientation of the curves is due to the computation of the energy ratio E_i based on the right ear (which is closer to the speech sound source; so it receives more speech energy).

Since we use a binary decision, we derive location dependent thresholds $T_i(\theta_1, \theta_2)$ across frequency channels that minimize the overall error rate. An error occurs when the decision differs from the ideal binary mask which is equivalent to $D_i > T_i$ and E_i <0.5, or $D_i < T_i$ and E_i >0.5 (where D_i refers to the peak location for low frequencies and to the computed IID for high frequencies). Minimum error rates range from 1.5% to 3.5% for low and high frequencies and up to 8% for middle frequencies.

Our model consists of a simple time-domain mechanism for speech segregation for a mixture of two sound sources. Assume that the desired source corresponds to location θ_1 ($\theta_1 > \theta_2$) and perfect localization has been achieved in a prior stage. Then, at every time frame *j* and for every frequency channel *i* we obtain a binary masking coefficient $\delta_{i,j}$ based on the following equation:

$$\delta_{i,j} = \begin{cases} 1, & \text{if } f_i \le 1.5\,k\text{Hz and } \tau_{\max}^i > T_i(\theta_1, \theta_2) \\ 1, & \text{if } f_i > 1.5\,k\text{Hz and } L_i > T_i(\theta_1, \theta_2) \\ 0, & \text{otherwise} \end{cases} \qquad (5)$$

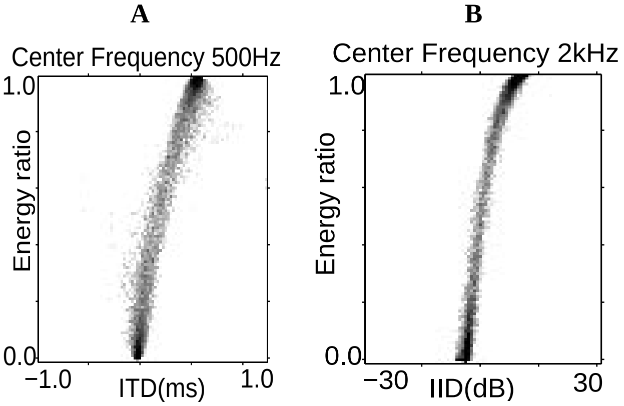

Figure 3: **A**: Relationship between ITD and the energy ratio for channel 40 (CF=500Hz). **B**: Relationship between IID and the energy ratio for channel 100 (CF=2kHz). Statistics are obtained for speech at 30° and interfering noise at -10°.

4 Results

Resynthesized signals have been extensively used to assess model evaluation. In our model we use a resynthesis method described by Brown and Cooke [10]. We compare our model with the Wang and Brown pitch-based model for speech segregation [3] across 10 types of noise interference. For the first criterion we measure independently the percentage of energy loss (EL) and the percentage of residual noise (RN) as defined below:

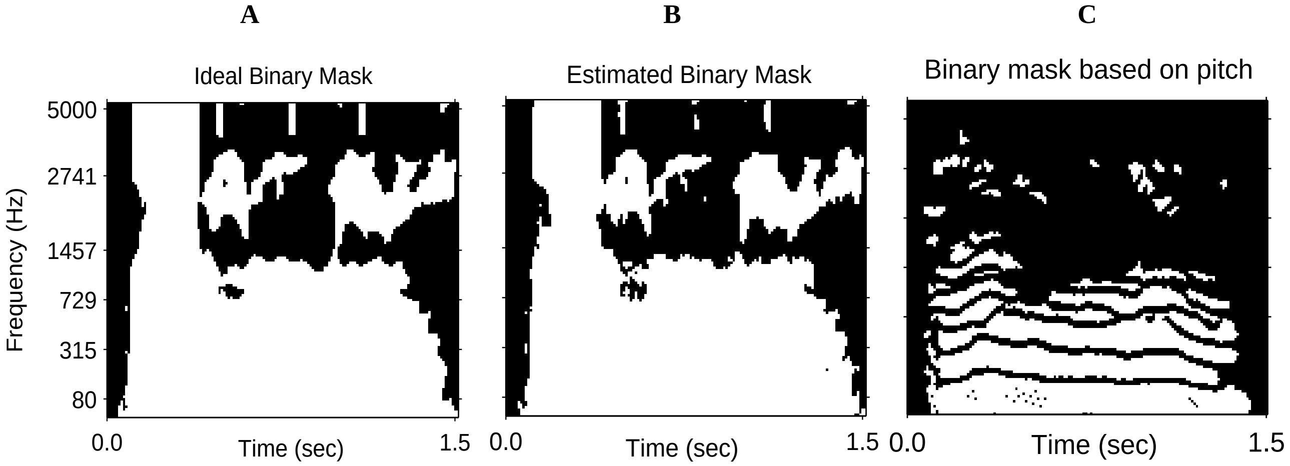

Figure 4: Results for a mixture of male speech at 30° and telephone ringing at -10°: the ideal binary mask (**A**), the estimated binary mask using our method of separation (**B**) and the Wang-Brown algorithm (**C**). Whiter regions correspond to the speech stream.

$$EL = \frac{\sum_t d_1^2(t)}{\sum_t s_I^2(t)} \times 100\% \qquad (6)$$

$$RN = \frac{\sum_t d_2^2(t)}{\sum_t s_e^2(t)} \times 100\% \qquad (7)$$

where the above are resynthesized signals using the following binary masks: $s_I(t)$ is obtained using the ideal binary mask, $d_1(t)$ using a binary mask that corresponds to regions selected in the ideal binary mask but missing from the estimated mask, $d_2(t)$ using a binary mask which corresponds to regions missing from the ideal mask but present in the estimated mask, and $s_e(t)$ using the estimated binary mask (Table 1). The second criterion represents the relative difference between $s_I(t)$ and $s_e(t)$ measured in decibels (Fig. 5):

$$D = 10\log_{10} \frac{\sum_t s_I^2(t)}{\sum_t (s_I(t) - s_e(t))^2} \qquad (8)$$

When compared with the Wang-Brown pitch-based algorithm, the estimated binary mask from our model recovers substantially more energy and constitutes a much better approximation to the ideal binary mask (Fig. 4). The energy loss (EL) decreases considerably in our model without increasing the residual noise (RN). At the same time, we observe a large increase for the second criterion, the relative difference from the ideal binary mask. Note, however, that the Wang-Brown model is a monaural system, where ours is binaural with two sensors.

By analyzing the statistics obtained in Sect. 3, we observe that the relationships between binaural cues and energy ratios are basically signal-independent. Hence, we expect a similar performance for our model when tested on a different corpus than the training one.

5 Discussion

In this paper, we have focused on the speech segregation problem using binaural cues (ITDs and IIDs). Our goal is to estimate an ideal binary mask that was psychoacoustically motivated by the auditory masking effect.

We have observed that there exists a relationship between ITDs and the *a priori* SNR in low frequency channels, as well as a relationship between IIDs and the *a priori* SNR in high frequency channels. We have analyzed these relationships across all frequency channels based on a corpus of 100 mixtures. Thresholds for binary decision rules are determined in order to minimize the overall error rate. When tested on the entire corpus, our algorithm approximates well the ideal binary mask and yields much improved performance over pitch-based monaural algorithms.

Acknowledgments. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-0027). We wish to thank G. Hu for many helpful discussions and assistance in preparation.

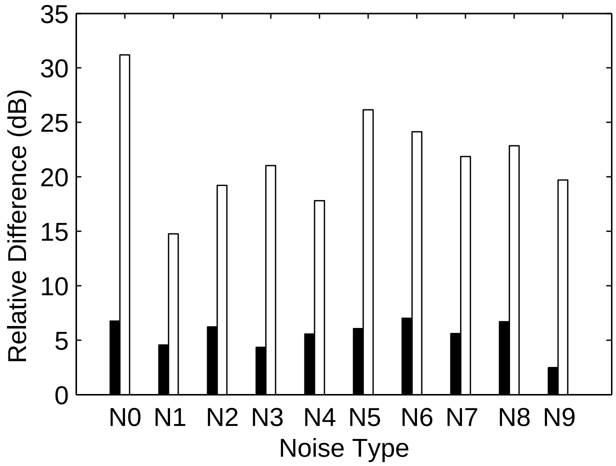

Figure 5: Relative difference from the ideal binary mask using pitch-based algorithm (black bar) and our model (white bar) for voiced speech mixed with ten different types of noise (N0=1kHz tone; N1=random noise; N2=noise burst; N3=”cocktail party”; N4=rock music; N5=siren; N6=trill telephone; N7=female speech; N8=male speech; N9=female speech). The voice source is located at 30° and the noise source at -10°.

Table 1: Percentage of energy loss (EL) and residual noise (RN) (same corpus as Fig. 5).

Noise Type	Pitch-based method		Our model	
	EL%	NI%	EL%	NI%
N0	22.80	0	0.04	0.02
N1	33.74	4.31	1.49	2.05
N2	20.81	4.50	0.04	1.28
N3	37.03	1.40	0.55	0.34
N4	26.74	3.16	1.09	0.79
N5	27.87	0.04	0.10	0.04
N6	21.23	0.41	0.22	0.20
N7	27.56	3.44	0.52	0.33
N8	22.47	0.57	0.15	0.52
N9	30.61	34.33	0.73	0.54
Average	27.09	5.29	0.49	0.61

References

[1] A. S. Bregman, *Auditory Scene Analysis*, Cambridge, MA: MIT press, 1990.

[2] J. Blauert, *Spatial Hearing - The Psychophysics of Human Sound Localization*, Cambridge, MA: MIT press, 1997.

[3] D. L. Wang and G. J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Trans. Neural Net.*, vol. 10, pp. 684-697, 1999.

[4] W. Lindemann, “Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation for lateralization for stationary signals,” *J. Acoust. Soc. Am.*, vol. 80, pp. 1608-1622, 1986.

[5] W. Gaik, “Combined evaluation of interaural time and intensity differences: Psychoacoustical results and computer modeling,” *J. Acoust. Soc. Am.*, vol. 94, pp. 98-110, 1993.

[6] M. Bodden, “Modeling human sound-source localization and the cocktail-party-effect,” *Acta Acoustica*, vol. 1, pp. 43-55, 1993.

[7] W. G. Gardner and K. D. Martin, “HRTF measurements of a KEMAR dummy-head microphone,” MIT Media Lab Perceptual Computing Technical Report #280, 1994.

[8] R. M. Stern and H. S. Colburn, “Theory of binaural interaction based on auditory-nerve data. IV. A model of subjective lateral position,” *J. Acoust. Soc. Am.*, vol. 64, pp. 127-140, 1978.

[9] R. F. Lyon, “A computational model of binaural localization and separation,” *Proceedings of IEEE ICASSP*, 1983.

[10] G. J. Brown and M. P. Cooke, “Computational auditory scene analysis,” *Computer Speech and Language*, vol. 8, pp. 297-336, 1994.

[11] M.P. Cooke, *Modeling Auditory Processing and Organization,* Cambridge, U.K.: Cambridge University Press, 1993.

[12] M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, “Robust automatic speech recognition with missing and unreliable acoustic data,” *Speech Comm.*, vol. 34, pp. 267-285, 2001.

[13] L. A. Jeffress, “A place theory of sound localization,” *J. Comp. Physiol. Psychol*., vol. 41, pp. 35-39, 1948.

[14] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th Edition, San Diego, CA: Academic Press, 1997.

[15] B. C. J. Moore, B. R. Glasberg and T. Baer, “A model for the prediction of thresholds, loudness and partial loudness,” *J. Audio Eng. Soc.*, vol. 45, pp. 224-240, 1997.

[16] L. Rayleigh, “On our perception of sound direction,” *Phil. Mag.*, vol. 13, pp. 214-232, 1907.