Technical Report OSU-CISRC-6/02-TR16
Department of Computer and Information Science
The Ohio State University
Columbus, OH 43210-1277

Ftp site: **ftp.cis.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2002**
File: **TR16.pdf**
Web site: **http://www.cis.ohio-state.edu/research/tech-report.html**

Speech Segregation Based on Sound Localization

Nicoleta Roman
Department of Computer and Information Science
The Ohio State University, Columbus, OH 43210, USA
niki@cis.ohio-state.edu

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

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

Correspondence should be directed to N. Roman: Department of Computer and Information Science, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210. Phone: (614)-292-7402, URL: www.cis.ohio-state.edu/~niki.

ABSTRACT

At a cocktail party, we can selectively attend to a single voice and filter out all the other acoustical interferences. How to simulate this perceptual ability remains a great challenge. This paper describes a novel machine learning approach to speech segregation, in which a target speech signal is separated from interfering sounds using spatial location cues: interaural time differences (ITD) and interaural intensity differences (IID). The auditory masking effect motivates the notion of an "ideal" time-frequency binary mask, which selects the target if it is stronger than the interference in a local time-frequency (T-F) unit. We observe that within a narrow frequency band, modifications to the relative strength of the target source with respect to the interference trigger systematic deviations for ITD and IID. For a given spatial configuration, this interaction produces characteristic clustering in the binaural feature space. Consequently, we perform pattern classification in order to estimate ideal binary masks. A systematic evaluation shows that the resulting system produces masks very close to ideal binary ones, and gives a significant improvement in performance over an existing approach, as quantified by changes in signal-to-noise ratio before and after segregation.

INTRODUCTION

The perceptual ability to detect, discriminate and recognize one utterance in a background of acoustic interference has been studied extensively under both monaural and binaural conditions (Bregman, 1990; Blauert, 1997; Bronkhorst, 2000). The human auditory system is able to segregate a speech signal from an acoustic mixture using various cues, including fundamental frequency (F0), onset time and location, in a process that is known as *auditory scene analysis* (ASA) (Bregman, 1990). F0 is widely used in computational ASA systems that operate upon monaural input – however, systems that employ only this cue are limited to voiced speech (Brown and Cooke, 1994; Wang and Brown, 1999). On the other hand, location (binaural) cues have the advantage of being generally independent of the signal content and can be used to track a sequence of voiced and unvoiced components that originate from the same location in space.

It is widely acknowledged that for human audition, interaural time differences (ITD) are the main localization cue used at low frequencies (<1.5 kHz), whereas in the high-frequency range both interaural intensity differences (IID) and interaural time differences between the envelopes of the signals (IED) are used (Blauert, 1997). The resolution of the binaural cues has implications in both localization and recognition tasks. Experiments show that listeners can reliably detect 10-15 μs ITDs from the median plane, which correspond to a difference in azimuth of between 1 and 5 degrees. On the other hand, the smallest detectable change in IID by the human auditory system is about 0.5 dB to 1 dB at all frequencies. Resolution deteriorates as the reference azimuth gets larger, and the difference limen for ITDs can be as much as 10 degrees when the reference source is located far to the side of the head (Blauert, 1997).

Classical models for processing binaural cues compare the acoustic signals at the two ears, although they explain the binaural interaction through different mechanisms. These include extensions of the Jeffress coincidence model (Jeffress, 1948; Lindemann 1986; Gaik 1993), the equalization and cancelation (EC) theory (Durlach, 1972) and auditory nerve based models (Colburn, 1973). The goal of this line of research is to explain experimental data for a number of psychoacoustical phenomena including lateralization, binaural masking levels, and the precedence effect (for a review see Stern and Trahiotis, 1995).

Increased speech intelligibility in binaural listening compared to the monaural case has also prompted research in designing cocktail-party processors based on psychoacoustic principles (Lyon, 1983; Slatky, 1993; Bodden, 1993; Liu *et al.*, 2001). Most cocktail-party-processor designs utilize the following observation: as the relative strength of the interference with respect to the target increases, certain attributes of the auditory event including location and extent change systematically compared to the case of the target source alone. In particular, building on a previous cross-correlation model for sound localization, Bodden (1993) proposed a model that estimates optimal time-varying Wiener coefficients for all critical bands by comparing the neural excitation patterns in cross-correlation with stored patterns obtained from clean speech. Although computationally expensive, Bodden's model has shown that psychoacoustically motivated auditory mechanisms can produce substantial enhancement in speech intelligibility (Bodden, 1996).

In this study, we propose a sound segregation model using binaural cues extracted from the responses of a KEMAR dummy head that realistically simulates the filtering process of the head, torso and external ear (Burkhard and Sachs, 1975). A typical approach for signal reconstruction uses a time-frequency (T-F) mask: T-F units are weighted selectively in order to enhance the target signal. We employ an *apriori* ideal binary mask that is motivated by the human auditory masking phenomenon, in which a stronger signal masks a weaker one in the same critical band (Moore, 1997). If the original unmixed signals are available, one can construct this ideal mask in the following way: retain the T-F units for which target energy exceeds interference energy and discard the other units. Ideal masks generate high quality reconstruction for a variety of signals, and similar binary masks have been shown to provide a very effective front-end to robust speech recognition (Cooke *et al.*, 2001). Hence, our model aims to estimate an ideal binary mask using information about the spatial configuration.

Statistics for the relationship of the relative strength between sources and the deviation of the binaural cues are at the core of our system. We show for mixtures of multiple sound sources that there exists a strong correlation between the relative strength and ITD/IID, resulting in a characteristic clustering across frequency bands. Our aim is to maximize the performance of the system independently for different spatial configurations. Consequently, we employ a nonparametric classification method to determine decision regions in the ITD-IID feature space that correspond to an optimal estimate for the ideal mask. We systematically evaluate the system for configurations of two sound sources in which the target position moves from the median plane to the side of the head and the smallest separation from the interfering source is 5 degrees. We also show that the performance of the model for more than two sources is comparable to the results from the ideal binary mask, although as expected the overall signal-to-noise ratio (SNR) drops.

The rest of the paper is organized as follows: the next section contains an overview of the model. Section II describes the peripheral auditory model. Section III describes the azimuth

localization algorithm. Section IV is mainly devoted to the ideal binary mask estimation, which constitutes the core of the model. Section V presents the performance of the system and a quantitative comparison with the Bodden model. In the last section we give further discussions and future directions.

I. MODEL ARCHITECTURE

Our model consists of the following four stages: 1) a model of the auditory periphery; 2) binaural cue extraction and azimuth localization for both target and interference based on a cross-correlation mechanism; 3) estimation of the ideal binary mask; and 4) reconstruction of the target signal. Figure 1 illustrates the model architecture for the case of two sound sources.

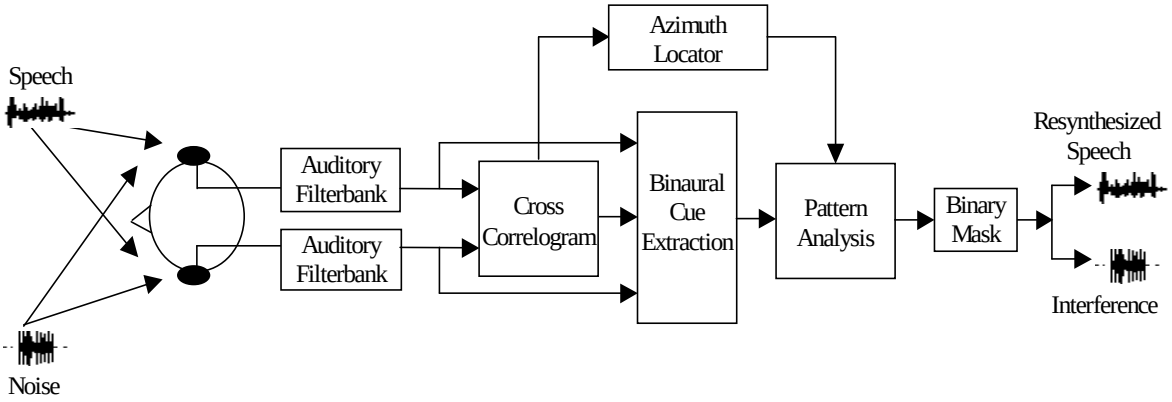

Figure 1. Schematic diagram of the model. Binaural signals are obtained by convolving input signals with head related impulse responses (HRIR). A model of the auditory periphery is employed. Azimuth localization for all the sources is based on a cross-correlation mechanism. ITD and IID are computed independently for different frequency channels. A pattern analysis block produces an estimation of an ideal binary mask, which enables the reconstruction of the target signal and the interfering sound.

The input to our model is a mixture of two or more signals presented at different, but fixed, locations: target speech and acoustic interference, sampled at 44.1 kHz. The sources are assumed to be in the horizontal plane, therefore only azimuth localization is considered here. We follow a standard procedure for simulating free-field acoustic signals from monaural signals (no reverberations or echoes are considered). Binaural signals are obtained by filtering monaural signals with head-related transfer functions (HRTF) corresponding to the direction of incidence. The responses to multiple sources are added at each ear. HRTFs introduce a natural combination of ITD and IID into the signals that is extracted by subsequent stages of our model. We utilize here a catalogue of HRTF measurements collected by Gardner and Martin (1994) from a KEMAR dummy head under anechoic conditions.

The auditory periphery is simulated using a filterbank that models the cochlear filtering mechanism. In addition, the gains of the filters are adjusted to account for middle ear transfer, which is direction-independent. The output of each filter is processed using a simple model for hair cell transduction, giving a representation of auditory nerve activity.

The auditory nerve responses from both ears are evaluated independently for all frequency bands in order to extract interaural differences. The most common method to determine ITD is cross-correlation of the corresponding left and right signals within individual frequency bands, which is calculated for time lags equally distributed in the plausible range (–1 ms to 1 ms). For azimuth localization we use only the information derived from ITD. Due to some diffraction effects, a frequency dependent nonlinear transformation from the time lag axis to the azimuth axis is necessary. The set of cross-correlations for all frequency bands and at all times results in a 3D structure called the "cross-correlogram"– where the coordinates are given by frequency, azimuth, time. A cross-correlogram is further evaluated to extract spatial information. Assuming fixed sources, the source locations are obtained as the positions of the maxima in a pooled cross-correlogram (Shackelton *et al.*, 1992) – obtained by integrating the cross-correlogram across time and frequency. Further stages of our model use this spatial information: the number of sources, their locations and the target source location.

At the core of our system are decision rules that determine whether the target source is stronger than the interference in individual T-F units. The system is based on observed characteristic clustering of extracted ITD and IID features. The novelty of our approach lies in the introduction of independent learning for different spatial configurations and across all frequency bands in a joint ITD-IID feature space. For a given frequency channel and a stimulus configuration, conditional probabilities are estimated from samples of ITD, IID and the corresponding relative strength based on a corpus of training data. Therefore, auditory grouping is implemented based on proximity in the ITD-IID space. The output of this pattern analysis is a time-frequency mask, which is an estimate of the ideal binary mask. The time-frequency resolution for the current implementation is 20-ms time frames with 10 ms overlapping between consecutive time frames, and 128 frequency channels that cover the range of 80 Hz to 5 kHz.

The last stage of the model is the reconstruction path, which allows the target signal to be recovered from the target signal from the acoustic mixture by masking the T-F units dominated by interference.

II. AUDITORY PERIPHERY

It is widely acknowledged that cochlear filtering can be modeled by a bandpass filterbank. The filterbank employed here consists of 128 fourth-order gammatone filters (Patterson *et al.*, 1988) following an implementation by Cooke (1993). The impulse response of the *i*th filter has the following form:

$$g_i(t) = \begin{cases} t^3 \exp(-2\pi b_i t)\cos(2\pi f_i t + \phi_i), & \text{if } t \geq 0 \\ 0, & \text{else} \end{cases} \quad (1)$$

where b_i is the decay rate of the impulse response, related to the bandwidth of the filter, f_i is the center frequency of the filter, and ϕ_i is the phase (here we set ϕ_i to zero).

The equivalent rectangular bandwidth (ERB) scale is a psychoacoustic measure for the auditory filter bandwidth at each frequency along the cochlea. The center frequencies f_i are

equally distributed on the ERB scale between 80 Hz and 5 kHz, and specifically for each filter we set the bandwidth according to the following equations (Glasberg and Moore, 1990):

$$ERB(f) = 24.7(4.37f/1000 + 1) \quad (2)$$

$$b_i = 1.019ERB(f_i) \quad (3)$$

Since the HRTF reflects the filtering effects due to pinna and meatus but not the middle ear we adjust the gains of the gammatone filters in order to simulate the middle ear transfer function; such data is provided by Moore *et al.* (1997). In the final step of the peripheral model, the output of each gammatone filter is half-wave rectified in order to simulate firing rates of the auditory nerve. Saturation effects are modeled by taking the square root of the signal.

Psychophysical models for sound localization generally employ envelopes of the responses in the high-frequency range. Therefore, we additionally extract envelopes using the Hilbert transform for channels with center frequency above 1.5 kHz. In Section IV we present a discriminability comparison of different binaural cues.

III. AZIMUTH LOCALIZATION

Current models of azimuth localization almost invariably start with the cross-correlation mechanism proposed by Jeffress. Cross-correlation provides excellent time delay estimation for broadband stimuli, and for narrowband stimuli in the low-frequency range. However, for high-frequency narrowband signals it produces multiple ambiguous peaks. Here we use the normalized cross-correlation computed at lags equally distributed in the plausible range from –1 ms to 1 ms $(-44 < \tau < 44)$ using an integration window of 20 ms (K=880). The cross-correlation is computed for all frequency channels and updated every 10 ms, according to the following formula for frequency channel i, time frame j and lag τ :

$$C(i,j,\tau) = \frac{\sum_{k=0}^{K-1}(l_i(j-k)-\bar{l}_i)(r_i(j-k-\tau)-\bar{r}_i)}{\sqrt{\sum_{k=0}^{K-1}(l_i(j-k)-\bar{l}_i)^2}\sqrt{\sum_{k=0}^{K-1}(r_i(j-k-\tau)-\bar{r}_i)^2}} \quad (4)$$

where l_i, r_i refer to the left and right auditory periphery output of the ith channel, and $\bar{l}_i$, $\bar{r}_i$ refer to their mean values estimated over the integration window.

For each frequency channel, ITD is estimated as the lag corresponding to the position of maximum in the cross-correlation function. Diffraction effects introduce weak frequency dependences for ITD (MacPherson, 1991). As a result, we derive frequency-dependent nonlinear transformations to map the time-delay axis onto the azimuth axis, resulting in a cross-correlogram $C(i,j,\varphi)$ where φ denotes azimuth. Fig. 2A shows three ITD-azimuth mappings, for channels with center frequencies of 500 Hz, 1 kHz, 3 kHz. The functions are monotonic, being sigmoidal at low frequencies where diffraction effects are greater and increasingly linear at high frequencies.

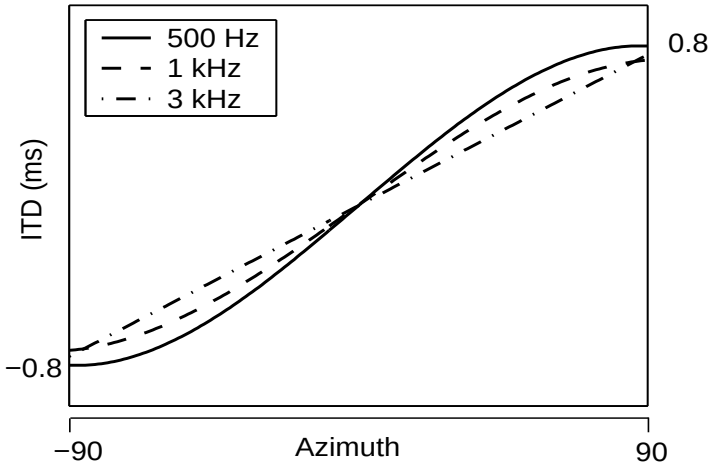

Figure 2. Functions relating azimuth to ITD for three auditory channels with center frequencies of 500 Hz, 1 kHz, and 3 kHz.

In addition, a 'skeleton' $S(i, j, \varphi)$ is formed by replacing the peaks in the cross-correlogram with gaussians whose widths are narrower than the original peaks. Here the width is linear with respect to the center frequency of the channel. This technique sharpens the cross-correlogram, an effect similar to the contralateral inhibition mechanism in the Bodden system.

The cross-correlation method provides inconsistent results when two acoustic sources are present. Figure 3 shows the cross-correlation functions (Fig. 3A) and the skeleton cross-correlogram (Fig. 3B) for a mixture of male speech presented at 30° and female speech presented at –10° at time frame 40 (i.e. 400 ms from the starting point). For frequency channels that are dominated by one source, activity is observed near the true location of that source. For T-F units where the two sources overlap the peak deviates, generally being closer to the more intense source. Peaks at both locations can occur in high-frequency channels – this ambiguity is due to the periodicity of the cross-correlation function. Hence, if little overlapping occurs for a sufficient number of channels a good estimate of the two source locations can be obtained at every time frame by pooling the cross-correlogram across all frequency channels. At time frame j and azimuth φ, this yields the following pooled cross-correlogram:

$$p(j, \varphi) = \sum_i S(i, j, \varphi) \tag{5}$$

Improved localization results are obtained using the skeleton cross-correlogram proposed here over the standard cross-correlation, where summing across frequencies produces sharper peaks for the two locations (compare the bottom plots in Fig. 3A and Fig. 3B). In Fig. 3C we display the pooled cross-correlogram for a signal of duration 150 frames (i.e. 1.5 seconds). Peaks in the pooled cross-correlogram indicate the locations of active sources at every frame. Assuming fixed sources, multiple locations can be reliably determined by further summating the pooled cross-correlogram across time as shown in the bottom plot of Fig. 3C. This represents our method for azimuth localization.

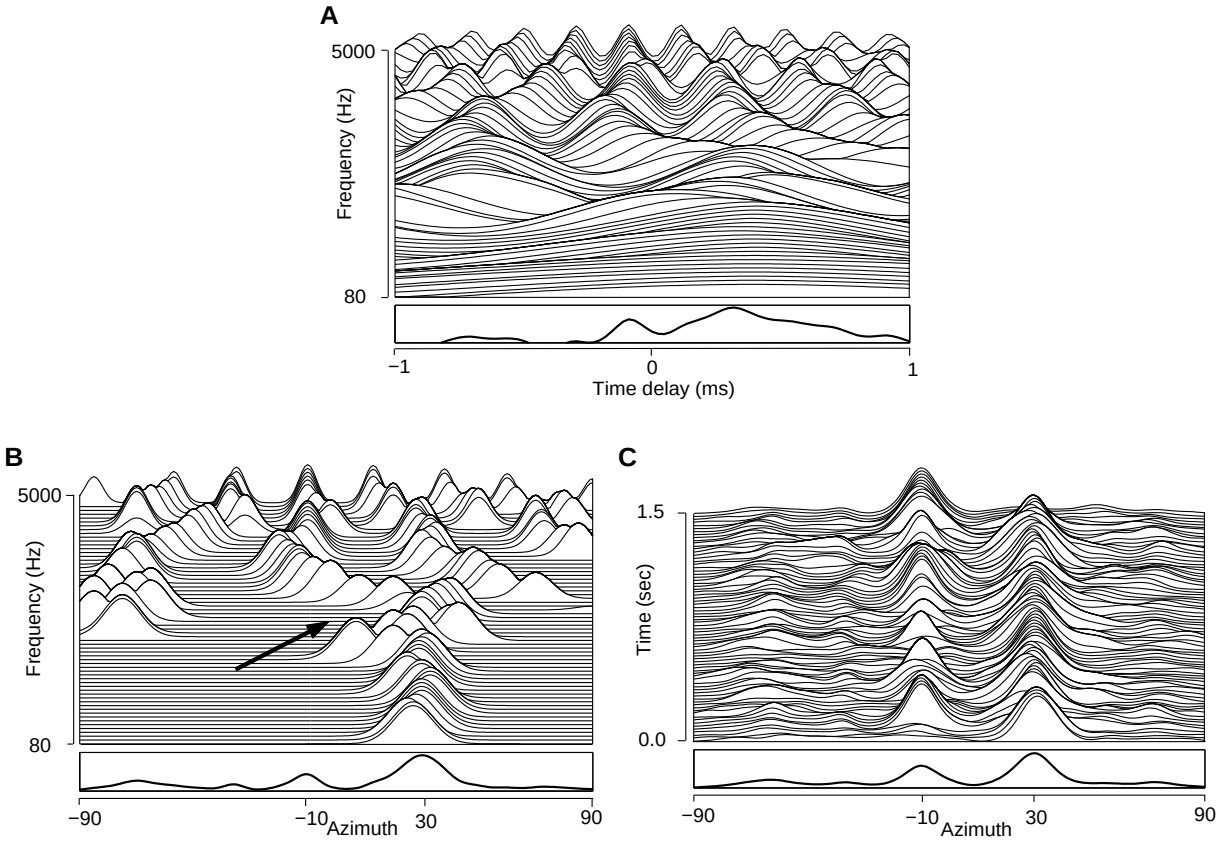

Figure 3. Azimuth localization for a mixture of male utterance at 30° and female utterance at -10°. The bottom plot in each panel shows a summation across all rows. **A**: Cross-correlation functions for 128 frequency channels in the range 80 Hz – 5 kHz at time frame 40 (i.e. 400 ms after the start of the stimulus). For clarity, only every other channel is shown, resulting in 64 channels. **B:** Skeleton cross-correlogram for the same time frame. The arrow indicates channels that contain roughly equal energy from both target and interference. **C**: Pooled cross-correlogram for a stimulus of duration 1.5 seconds, shown every 20 ms.

IV. IDEAL MASK ESTIMATION

The objective of this stage of the model is to develop an efficient mechanism for estimating the ideal binary mask. We propose an estimation method based on the following observation regarding the auditory interaction of multiple sources. In a narrow band, the ITD and IID corresponding to the target source exhibit azimuth-dependent characteristic values. As the interference from additional sound sources increases, ITD and IID systematically shift away from these values. Consequently, in a local T-F unit both binaural cues can be potentially used to determine whether the target signal dominates.

In what follows, we analyze this phenomenon for the case of pure tones. This analysis serves to motivate our proposed algorithm.

A. Pure Tones

We consider two sources emitting pure tones in a narrow band. In this case, the left-ear and the right-ear responses are given by:

$$\begin{cases} l(t) = \left|H_1^l(\omega_1)\right| A_1 \sin(\omega_1 t) + \left|H_2^l(\omega_2)\right| A_2 \sin(\omega_2 t + \Delta\varphi) \\ r(t) = \left|H_1^r(\omega_1)\right| A_1 \sin(\omega_1 t + \omega_1 d_1) + \left|H_2^r(\omega_2)\right| A_2 \sin(\omega_2 t + \omega_2 d_2 + \Delta\varphi) \end{cases} \tag{6}$$

where A_i is the amplitude, ω_i is the frequency, d_i corresponds to the interaural time delay, and $H_i^r(\omega_i)$ and $H_i^l(\omega_i)$ represent respectively the right and left HRTF, for the *i*th source. $\Delta\varphi$ is the sum of phase differences between the initial signals and those due to the arrival times of the signals at the left ear.

To simplify, we neglect the magnitude of the HRTF response in analyzing ITD, which represents a reasonable assumption in a narrowband low-frequency range. The cross-correlation function for infinite-duration signals is obtained by:

$$c(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} r(t) l(t+\tau) dt \tag{7}$$

Observe that in approximating the cross-correlation function in a finite duration, there exists a tradeoff between the difference in frequency $|\omega_1 - \omega_2|$ and the total integration time. Therefore, we study the cross-correlation under the following two conditions:

Case 1*:* $\omega_1 = \omega_2 = \omega$

In this case, we have:

$$\begin{aligned} c(\tau) = {} & \frac{A_1^2}{2}\cos(\omega(\tau - d_1)) + \frac{A_2^2}{2}\cos(\omega(\tau - d_2)) + \\ & + A_1 A_2 \cos\left(\omega\left(\tau - \frac{d_1 + d_2}{2}\right)\right) \cdot \cos\left(\Delta\varphi + \omega\frac{d_2 - d_1}{2}\right) \end{aligned} \tag{8}$$

Due to the periodicity of $c(\tau)$, we study the cross-correlation function on a 2π interval centered at $\omega(d_1 + d_2)/2$. Without loss of generality, assume that the phase differences $\omega d_1, \omega d_2$ are in this interval; otherwise, simply shift the phases with multiples of 2π. To fix the discussion let $d_1 < d_2$. By observing the deviation of the peak location $\tau_{\max}$ from the middle of the two sources, $(d_1 + d_2)/2$, we obtain the stronger source:

$$\tau_{\max} > (d_1 + d_2)/2 \Leftrightarrow A_1 < A_2 \tag{9}$$

This result gives a threshold to decide which source is stronger based on ITD. Furthermore, we want to study how ITD changes with the relative strength $R = \frac{A_2}{A_1 + A_2} \in [0, 1]$. Variations in $\Delta\varphi$ affect the distribution of $\tau_{\max}$, and we thus carry out a probabilistic study on the relationship between ITD and *R*. A reasonable assumption is that $\Delta\varphi$ is uniformly distributed in the range $[-\pi, \pi]$. To simplify, we use the following notations: $\alpha = \omega\left(\tau_{\max} - \frac{d_1 + d_2}{2}\right) \in [-\pi, \pi]$ and $\beta = \omega\frac{d_2 - d_1}{2} \in [0, \pi]$, and derive the solution for α as follows:

$$\alpha = \arctan\left[\frac{(A_2^2 - A_1^2)\sin\beta}{(A_1^2 + A_2^2)\cos\beta + 2A_1A_2\cos(\Delta\varphi + \beta)}\right] + k\pi \tag{10}$$

where *k* is an integer. Fig. 4 shows the domain of $\alpha \in [-\pi, \pi]$; here the phases $-\beta$ and β correspond to the two source locations d_1 and d_2. The relation obtained in (9) restricts α to the interval $[0, \pi]$ for $A_1 < A_2$, and $[-\pi, 0]$ for $A_1 > A_2$. Hence, $k \in \{0, \pm 1\}$ is uniquely determined. For a continuous random variable X with distribution $p(x)$ and a differentiable function $g(x)$, the distribution of the variable $Y=g(X)$ can be obtained from $p(y) = \sum_{i=1}^{n} p(x_i) / |g'(x_i)|$, where x_i is a root for $y=g(x)$. Straightforward calculations result in the following formula for the distribution of α:

$$\begin{aligned} p(\alpha) &= \left|\frac{(A_2^2 - A_1^2)\sin\beta}{2\pi A_1A_2\sin^2\alpha}\right| \frac{1}{\sqrt{1 - \cos^2(\Delta\varphi + \beta)}} \\ &= \frac{|(A_2^2 - A_1^2)\sin\beta|}{\pi|\sin\alpha|\sqrt{4A_1^2A_2^2\sin^2\alpha - (A_2^2\sin(\alpha - \beta) + A_1^2\sin(\alpha + \beta))^2}}, \quad \alpha \in [\alpha_1, \alpha_2] \end{aligned} \tag{11}$$

where the bounds $\alpha_{1,2}$ are obtained from (10) by setting $\cos(\Delta\varphi + \beta) = \pm 1$.

For $R \to 1$, both bounds $\alpha_{1,2}$ converge to the same location β ($-\beta$ for $R \to 0$). Therefore, the probability distribution has a sharp peak indicating the time delay of the predominant source. As the difference in amplitude gets smaller (i.e. $R \to 0.5$), $\alpha_2 - \alpha_1 \to \pi$, which increases the uncertainty (variance) and spreads out the peak in the probability distribution. The distribution $p(\alpha)$ has two sharp peaks, i.e. $p(\alpha) \to \infty$ at the bounds $\alpha_{1,2}$ (where the denominator cancels). In order to study the trend of the peak location as the relative strength changes, we analyze the mean of the distribution, $\bar{\alpha}$, obtained by integrating α from (10) over $\Delta\varphi$ from $-\pi$ to π. Note that when the denominator in (10) cancels the integral must be decomposed for the corresponding continuous intervals and *k* modified accordingly. It can be shown that the limit of

$\bar{\alpha}$ as $R \to 1$ is β and $-\beta$ for $R \to 0$. In addition, the left and right limit as $R \to 0.5$ are $-\beta$ and β respectively. Moreover, simulations show that a good approximation is given by:

$$\bar{\alpha} \approx \begin{cases} \beta, & R > 0.5 \\ 0, & R = 0.5 \\ -\beta, & R < 0.5 \end{cases} \tag{12}$$

Case 2: $\omega_1 \neq \omega_2$

In this case, due to the orthogonality of sine waves of different frequencies the cross-correlation function becomes:

$$c(\tau) = \frac{A_1^2}{2}\cos(\omega_1(\tau - d_1)) + \frac{A_2^2}{2}\cos(\omega_2(\tau - d_2)) \tag{13}$$

A closed form solution for the peak location in this case does not exist. Instead, we analyze the behavior of the peak location for relatively close angles, i.e. $|\omega_1 d_1 - \omega_2 d_2| < \pi/2$. In this interval, we apply a second-order Taylor expansion as an approximation for the cosine, resulting in a simple solution: $\tau_{\max} = \frac{A_1^2\omega_1^2 d_1 + A_2^2\omega_2^2 d_2}{A_1^2\omega_1^2 + A_2^2\omega_2^2}$. Note that this is a monotonic function with respect to the relative strength *R*.

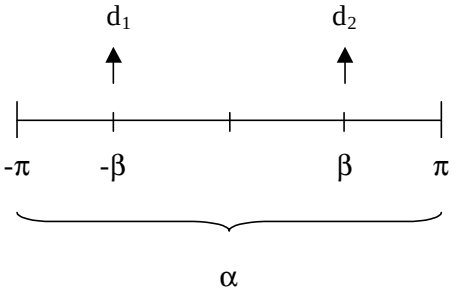

Figure 4. The range of variable α. Here d_1 and d_2 indicate the two source locations.

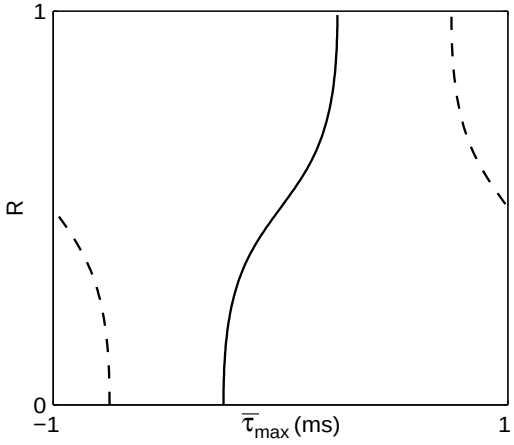

Figure 5. Theoretical approximation for the mean ITD, $\bar{\tau}_{\max}$, for two pure tones randomly distributed in a narrow band centered at 500 Hz. The y-axis corresponds to the relative strength R. Two cases are shown: $\beta = \pi/4$ (solid line) and $\beta = 3\pi/4$ (dashed line).

For the general case, we observe that as the frequencies ω_1 and ω_2 vary uniformly in a narrowband centered at ω, a good approximation for the mean of $\tau_{\max}$ is given by:

$$\bar{\tau}_{\max} = \frac{1}{\omega}\left(\frac{d_1 + d_2}{2} + \arctan\left[\frac{(A_2^2 - A_1^2)}{(A_1^2 + A_2^2)} \tan\beta \right] + k\pi \right), \quad k \in \{0, \pm 1\} \tag{14}$$

which is the solution for the maximum position in (13) when $\omega_1 = \omega_2$. This function is monotonically increasing with respect to *R* when $\beta < \pi/2$ and decreasing when $\beta > \pi/2$ (see Fig. 4). Fig. 5 shows the results when $\omega = 500$ Hz and β equals $\pi/4$ and $3\pi/4$, respectively.

A systematic change in *R* also results in a corresponding shift in IID. A similar discussion applies here. That is, the frequency difference between the two tones affects the spread of IID distribution. We do not study the case $\omega_1 = \omega_2$ since the results for IID distribution are complicated. In addition, IID is most reliable at high frequencies where filter bandwidths are large. Therefore, we consider the case $\omega_1 \neq \omega_2$. IID is approximated as the ratio of signal power at the two ears, resulting in the following expression:

$$\text{IID} = 10\log_{10} \frac{A_1^2 \mid H_1^r(\omega_1) \mid^2 + A_2^2 \mid H_2^r(\omega_2) \mid^2}{A_2^2 \mid H_1^l(\omega_1) \mid^2 + A_2^2 \mid H_2^l(\omega_2) \mid^2} \tag{15}$$

where the power of a signal $x(t)$ is $\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x^2(t)dt$. Note that IID is monotonic with respect to the relative strength *R*.

The above analysis suggests that the distribution of the binaural cues in a given filter channel is directly influenced by the filter bandwidth. To test this, we simulate left and right signals using Eq. 6, where the relative strength is fixed, $\Delta\varphi$ is uniformly distributed in the range $[-\pi, \pi]$ and $\omega_{1,2}$ in $[\omega - \Delta\omega, \omega + \Delta\omega]$. Figs. 6A and 6B show the mean and the variance of ITD as a function of *R* for the condition of $\omega = 500$ Hz, 30° azimuth separation, 20-ms integration time and four $\Delta\omega$ values in the range of 0 Hz to 200 Hz. In the figure, M_1 is the ITD mean as derived in (13) and it approximates well the case $\Delta\omega = 0$. M_2 is the ITD mean derived in (14) for the more general case $\Delta\omega \neq 0$. Similarly, Figs. 6C and 6D show results for IID when $\omega = 2.5$ kHz and five $\Delta\omega$ values in the range of 0 Hz to 400 Hz. Here, *M* is the IID mean as derived in (15). It is worth noting that the theoretical derivations of M_2 and *M* approximate well the simulation results when the bandwidth approaches the auditory filter ERB, which is 80 Hz for a 500 Hz center frequency and 300 Hz for 2.5 kHz. In addition, there is a systematic decrease in variance for both ITD and IID as $\Delta\omega$ approaches the ERB. This behavior generalizes to other frequencies as well.

To conclude, our analysis shows that ITD and IID undergo systematic shifts from the ideal target values as the relative strength *R* of two sinusoidal sources is changed. A comparison of the above theoretical derivations with the real data presented in the next subsection shows that the match is very close.

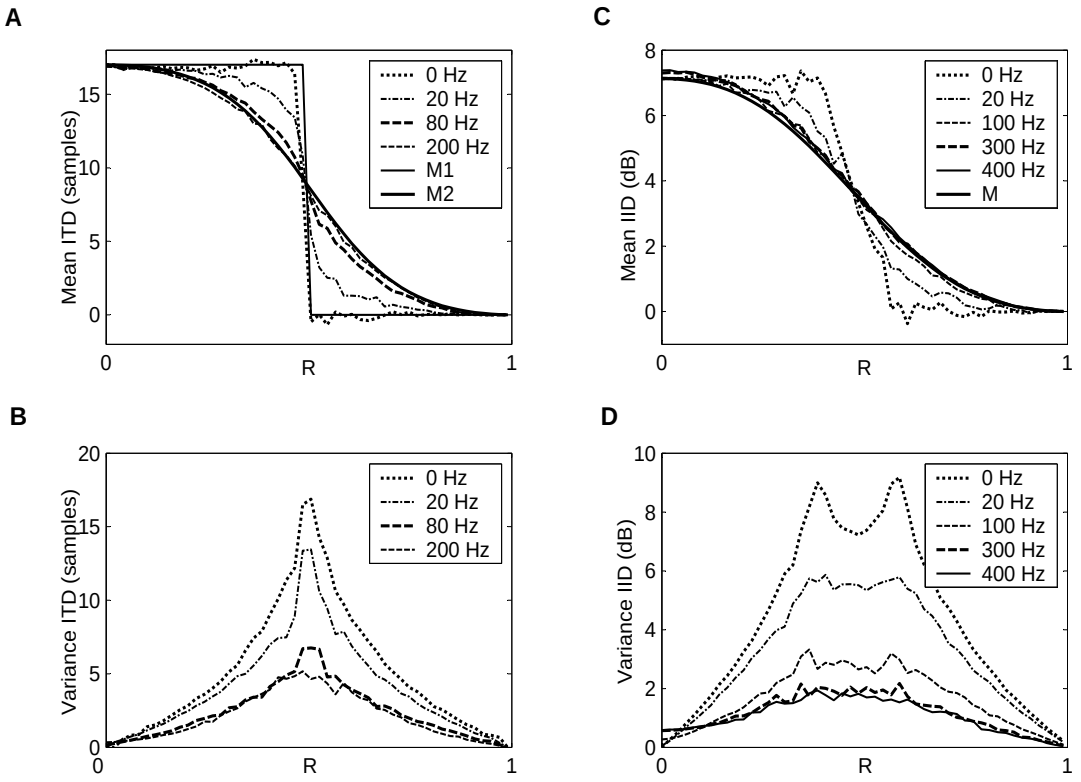

Figure 6. The influence of filter bandwidth on the mean and variance of ITD and IID with respect to the relative strength *R*. The data is from simulations of two pure tones uniformly distributed in a narrow band. One tone is at 0^{o} and the another is at 30^{o}. The sampling frequency is 44.1 kHz. **A**: Mean ITD as a function of *R* for 500 Hz center frequency and four bandwidths between 0 Hz and 200 Hz. The auditory filter ERB here is 80 Hz. M_1 and M_2 correspond to the theoretical mean ITD as derived in Eq. 13 and Eq. 14, respectively. **B**: ITD variance for the same condition as in A. **C**: Mean IID as a function of *R* for a 2.5 kHz center frequency and five bandwidths between 0 Hz and 400 Hz. *M* corresponds to the theoretical mean IID as derived in Eq. 15. The auditory filter ERB is 300 Hz. **D**: IID variance for the same condition as in C.

B. Model

The analysis of ITD and IID for pure tones shows relatively smooth changes with the relative strength *R* in narrow frequency bands. In order to capture this relationship in the context of real signals, statistics are collected for individual spatial configurations during training. We employ a training corpus consisting of 10 speech signals from the TIMIT database (Garofolo *et al.*, 1993): 5 male utterances and 5 female utterances as presented in Table I. The speaker ID in the table uniquely identifies the speaker in the TIMIT database where the first letter indicates the sex of the speaker. In the two-source case, we select S0-S4 to be target and the rest interference. In the three-source case, we have S0-S3 as target signals and the 2 interfering sets are S4-S6 and S7-S9.

TABLE I: SPEECH SIGNALS OF THE TRAINING SET

ID	Speaker ID	Utterance
S0	MKLS0	"Primitive tribes have an upbeat attitude"
S1	FCKE0	"Only the best players enjoy popularity"
S2	MCDC0	"Our aim must be to learn as much as we teach"
S3	FEAR0	"Development requires a long-term approach"
S4	FDMS0	"Poets, moreover, dwell on human passions"
S5	FETB0	"Change involves the displacement of form"
S6	FCMM0	"The system works as an impersonal mechanism"
S7	MJWS0	"Most assuredly ideas are invaluable"
S8	MRVG0	"False ideas surfeit another sector of our life"
S9	MJRH0	"But in every period it has been humanism"

Estimates for ITD, IID and R are extracted independently for all frequency channels. Computations are based on 20-ms time frames with 10-ms overlap between adjacent frames. Since the cross-correlation function is periodic, resulting in multiple peaks for mid to high frequencies, we consider the following strategy for estimating ITD. We study deviations from the target ITD for individual frequency channels, which is obtained from the ITD-azimuth mappings presented in Section III. Consequently, we compute ITD_i as the peak location of the cross-correlation function in the range $2\pi/\omega_i$ centered at the target ITD, where ω_i indicates the center frequency of the ith channel. IID_i corresponds to the mean power ratio at the two ears, expressed in decibels:

$$IID_i = 20\log_{10}\left(\sum_t r_i^2(t) \Big/ \sum_t l_i^2(t)\right) \tag{16}$$

where l_i and r_i refer to the left and right auditory periphery output of the ith channel, respectively. Note that in computing IID_i, we use 20 instead of 10 in order to compensate for the square root operation in the peripheral processing stage.

The relative amplitude is a measure of the relative strength between the target source and the acoustic interference, defined using root-mean-square values of the original signals at the "better" ear (the ear closer to the target source):

$$R_i = \sqrt{\sum_t s_i^2(t)} \Big/ \left(\sqrt{\sum_t s_i^2(t)} + \sqrt{\sum_t n_i^2(t)}\right) \tag{17}$$

where s_i refers to the response of the ith gammatone filter to the target signal and n_i the response to the acoustic interference (noise).

Fig. 7 shows empirical results obtained for a two-source configuration: target source in the median plane and interference at 30°. The scatter plot in Fig. 7A shows samples of ITD and *R* obtained for the channel with a center frequency of 500 Hz. In addition, we display the empirical mean of the samples and the theoretical one derived in (14). Similarly, Fig. 7B shows the results that describe the variation of IID with *R* for a channel with a center frequency of 2.5 kHz and compares the empirical mean with the one derived in (15). Note that R_i incorporates the HRTF responses at the better ear. Therefore, the *R* axis for the theoretical mean is converted accordingly. Fig. 7 exhibits a systematic shift of ITD and IID with respect to *R* for real signals. Moreover, the theoretical means obtained in the case of pure tones match the empirical ones very well. Similar matches are observed in other frequency channels and other spatial configurations.

The above observation extends to multiple-source scenarios. As an example, Fig. 8 displays smoothed histograms that show the relationship between *R* and both ITD (Fig. 8A) and IID (Fig. 8B) for a three-source situation. Samples correspond to a frequency channel with a center frequency close to 1.5 kHz for target at 0° (median plane) and two interferences at –30° and 30°. Note that the interfering sources introduce systematic deviations of the binaural cues. Consider a particularly troubling case: the target is silent and two interferences have equal energy in a given T-F unit. This results in binaural cues indicating an auditory event at half of the distance between the two interference locations; for our setup, it is 0° - the target location. However, the data in Fig. 8 suggest a low probability for this case. Fig. 8 instead shows a clustering phenomenon, suggesting that in most cases only one source dominates a T-F unit.

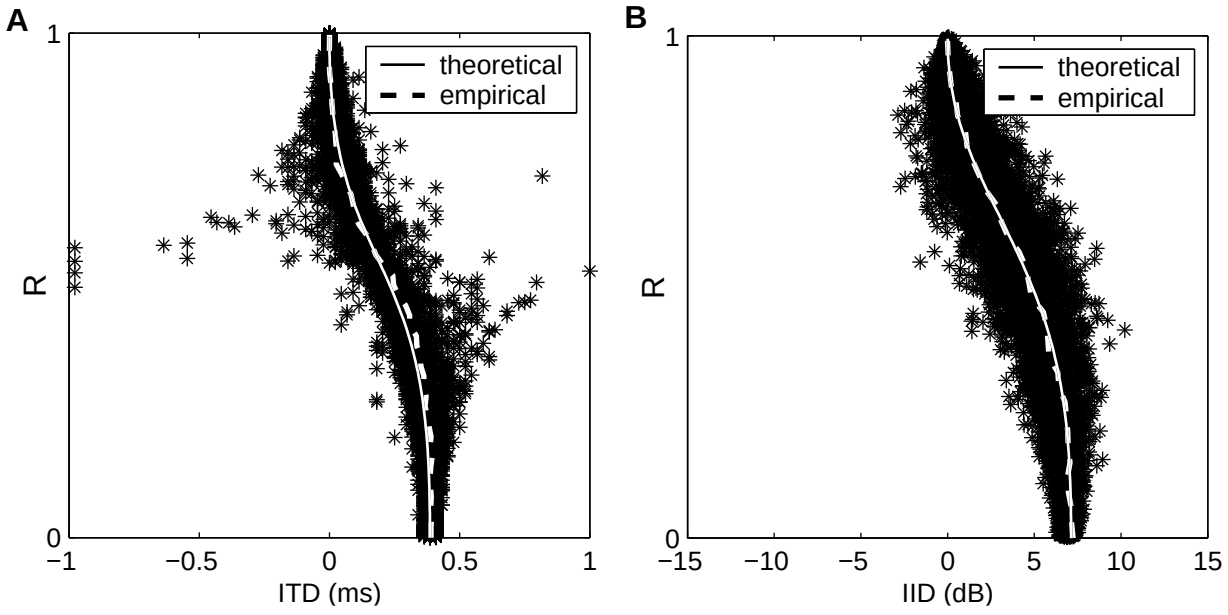

Figure 7. Relationship between ITD/IID and the relative strength *R* for a two-source configuration: target in the median plane and interference on the right side at 30°. **A**: The scatter plot shows ITD and R estimates from the training corpus for a channel with center frequency of 500 Hz. The solid curve shows the theoretical mean (see Eq. 14) and the dash curve shows the data mean. **B:** Results for IID for a filter channel with center frequency 2.5 kHz. The solid curve shows the theoretical mean (see Eq. 15) and the dash curve shows the data mean.

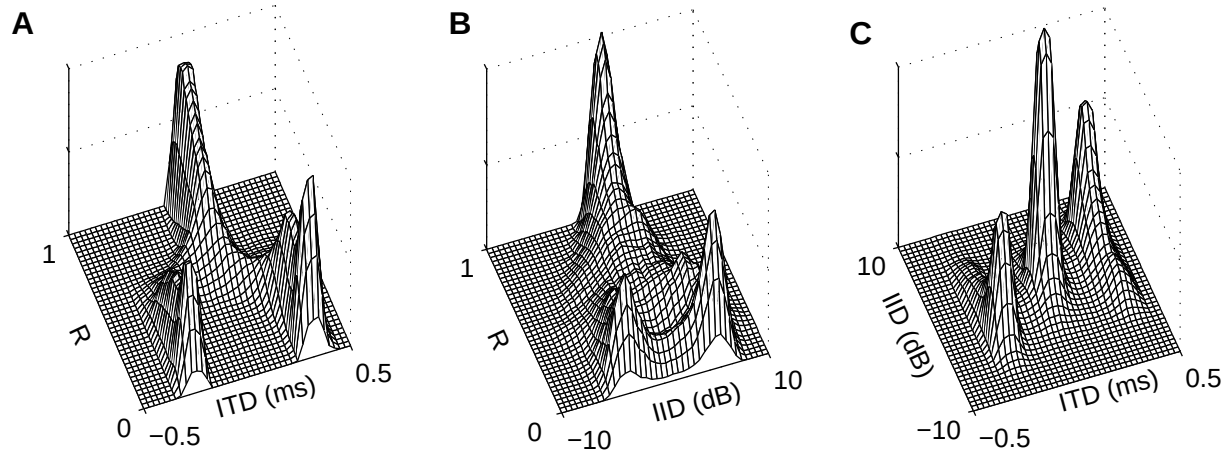

Figure 8. Relationship between ITD/IID and the relative strength *R* for a three-source configuration: target source in the median plane and interference at -30° and 30°. Statistics are obtained from the training corpus for a channel with center frequency close to 1.5 kHz. **A**: Histogram of ITD and R samples. **B**: Histogram of IID and *R* samples. **C**: Clustering in the ITD-IID space.

By displaying the information in the joint ITD-IID space, we observe a location-based clustering of the binaural cues, which is clearly marked by strong peaks that correspond to distinct active sources as shown in Fig. 8C. There exists a tradeoff between ITD and IID across frequencies, where ITD is most salient at low frequencies and IID at high frequencies. But a fixed cutoff frequency that separates the effective use of ITD and IID does not exist for different spatial configurations (see Fig. 9 below). This motivates our choice of a joint ITD-IID feature space that optimizes the system performance across different configurations. Differential training seems necessary for different channels given that there exist variations of ITD and, especially, IID values with different center frequencies.

Since the goal is to estimate an ideal binary mask, we focus on detecting decision regions in the 2-dimensional ITD-IID feature space for individual frequency channels. Consequently, standard supervised learning techniques can be applied. For the *i*th channel, we test the following two hypotheses. The first one is H_1: target is dominant or $R_i > 0.5$, and the second one is H_2: interference is dominant or $R_i < 0.5$. Based on estimates of the bivariate densities $p(x \mid H_1)$ and $p(x \mid H_2)$ the classification is done in accordance with the *maximum a posteriori* (MAP) decision rule: $p(H_1)p(x \mid H_1) > p(H_2)p(x \mid H_2)$. There exist a plethora of techniques for probability density estimation ranging from parametric techniques (e.g. mixture of gaussians) to nonparametric ones (e.g. kernel density estimators). We initially tried the EM algorithm for learning Gaussian mixtures (Duda *et al.*, 2001), but this did not prove to be robust due to the following factors: (i) the true number of mixing components is usually unknown, and (ii) the algorithm is sensitive to parameter initialization. Even for the two-source scenario, the method of computing ITD for mid- to high-frequencies can result in irregular distributions for the H_2 hypothesis (two-peak distribution). In order to completely characterize the distribution of the data here we use the kernel density estimation method independently for all frequency channels.

While the kernel density estimation method is well documented in the literature (Silverman, 1986), we summarize its essence here. Generally, the multidimensional kernel density estimate for n observations $\mathbf{x}_1,..,\mathbf{x}_n$ of dimensionality d is given by the following formula:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{nh_1...h_d} \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right) \tag{18}$$

where $\mathbf{x} = (x_1,...,x_d)$ is a feature vector, x_{ij} is the jth element of $\mathbf{x}_i$, K is a Gaussian function, and h_j's are parameters called bandwidths that define the amount of smoothing for the empirical distribution. In our case, the ITD-IID feature space has dimensionality $d = 2$. The selection of the smoothing parameters is critical to the success of the estimation process: for too small values it approximates the data well but generalizes poorly and for too large values the structure of the data distribution disappears. One approach for finding optimal values is the least-squares cross-validation method (LSCV) (Silverman, 1986). We employ the LSCV method for high dimensions and the Gaussian kernel given by Sain *et al.* (1994) (p. 808). Optimal smoothing values are chosen as local minima in the range $[\, n^{-1/6}\sigma_i / 4,\ 3n^{-1/6}\sigma_i / 2\,]$, where σ_i represents the variance of the data set in the ith dimension and n is the size of sample data set.

The performance of the system is then measured on independent test corpuses for different spatial configurations (see Sect. V). For the two-source scenario, one test set is the corpus collected by Cooke (1993), chosen because it is commonly used in computational ASA studies (Brown and Cooke, 1994; Wang and Brown, 1999; Wu *et al.*, 2002). The corpus contains 10 voiced speech signals and 10 noise intrusions, encompassing a variety of common acoustic interferences such as telephone ringing, rock music, and other speech utterances. In addition, we employ a second corpus containing 10 normal speech utterances from the TIMIT database (see Table II) as target mixed with the 10 intrusions from the Cooke corpus. In the case of three sources, we use for testing the Cooke corpus: 5 speech signals form the target set and the other 5 form one interference source. The 10 intrusions then form the second interference source. Therefore, in this three-source corpus every mixture contains two utterances plus an additional intrusion.

One cue not employed in our model is the interaural envelope difference or IED. Auditory models generally use IED in the high-frequency range (see for example Bodden, 1993) since the auditory system becomes gradually insensitive to interaural phase differences above 1.5 kHz. We have compared the individual performance of the three binaural cues: ITD, IID and IED, on a 1-dimensional classification task based on the kernel density estimation method presented above. Fig. 9 shows the error rates with respect to frequency channel for the classification task on the Cooke corpus as the testing set, where we consider two cases: target source in the median plane and the acoustic interference at 5° (Fig. 9A) and 30° (Fig. 9B). For IED results are given for the frequency range of interest - above 1.5 kHz (i.e. channel number > 80). As the source separation increases, error rates for IED and IID improve. On the other hand, ITD loses discriminability for high-frequency channels where the multiple-peak problem results in the same ITD for both target and interference (Fig. 9B). As indicated in Fig. 9, we have found no benefit for using IED after incorporating ITD and IID. This is the reason that IED is not included in our model. Fig. 9 also

displays the corresponding error rate for our model that uses the joint ITD-IID space and it gives the best performance across different spatial configurations.

TABLE II: TARGET SIGNALS OF THE TESTING SET

ID	Speaker ID	Utterance
S0	MWSB0	"Bright sunshine shimmers on the ocean"
S1	MDCD0	"Challenge each general's intelligence"
S2	MDHS0	"The Thinker is a famous sculpture"
S3	MTAA0	"Only lawyers love millionaires"
S4	MRPC1	"Biblical scholars argue history"
S5	FPKT0	"They make us conformists look good"
S6	FJRE0	"Artificial intelligence is for real"
S7	FPAC0	"A good attitude is unbeatable"
S8	FREH0	"Too much curiosity can get you into trouble"
S9	FBCH0	"Clear pronunciation is appreciated"

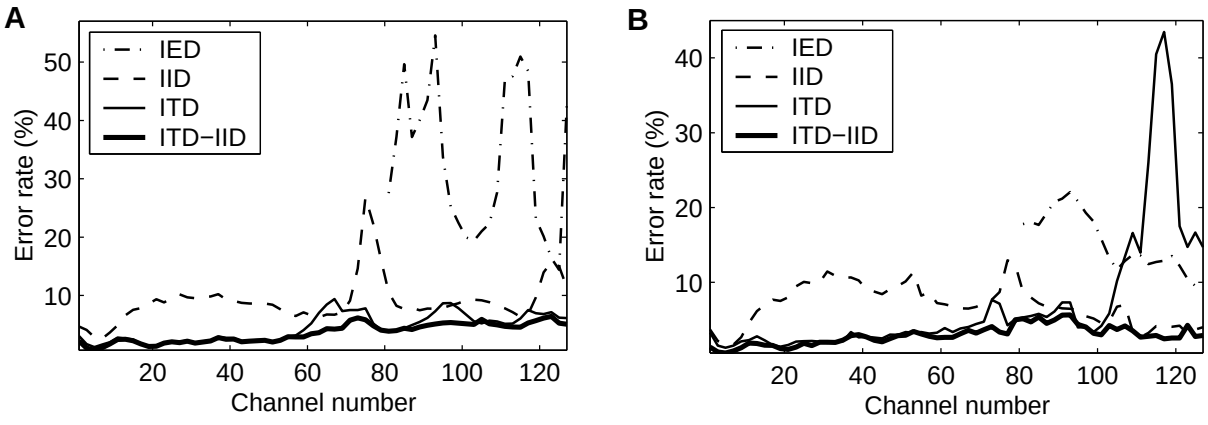

Figure 9. Discriminability comparison for the three binaural cues, ITD, IID and IED, and the joint ITD-IID space. Error rates are displayed as a function of channel number (frequency) for a classification task for two spatial configurations. **A**: Target source in the median plane and interference on the right side at 5°. **B**: Target source in the median plane and interference on the right side at 30°. IED results are shown for frequencies above 1.5 kHz, i.e. above channel number 80.

V. EVALUATION AND COMPARISON

A binary mask produced by the model described in the last section approximates very well the corresponding ideal binary mask. As an example, Fig. 10 shows a comparison between the ideal binary mask and the estimated mask for a mixture of target male speech presented at 0° and

interference female speech at 30° at the better ear. In the figure, a blank pixel indicates a T-F unit in which the target dominates. The two masks are very similar.

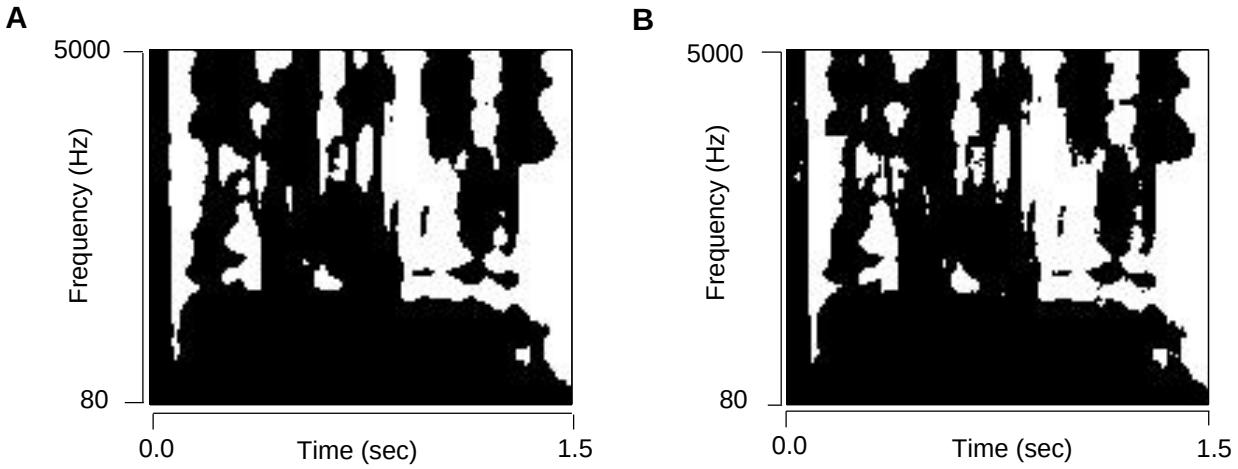

Figure 10. A comparison between an ideal binary mask (**A**) and the binary mask resulting from our model (**B**) for a mixture of male utterance in the median plane (target) and female utterance on the right side at 30° (interference). The black regions indicate those T-F units dominated by target speech.

In order to evaluate the performance of our system for speech segregation, a segregated signal is reconstructed from a binary mask following a method described by Brown and Cooke (1994). To quantitatively assess system performance, we measure in decibels the SNR using the original target speech before mixing:

$$SNR = 10\log_{10}\sum_t s_T^2(t) \Big/ \sum_t \left(s_T(t) - s_E(t)\right)^2 \tag{19}$$

where $s_T(t)$ represents the original target signal and $s_E(t)$ the estimated target reconstructed from the binary mask. One can similarly measure the SNR of the original mixture by replacing the denominator with $s_N(t)$, the noise intrusion. To avoid distortions, $s_T(t)$ and $s_N(t)$ represent the reconstructed signals using an all-one binary mask with original target and original intrusion as input, respectively. To minimize the loss of target energy we take advantage of the higher initial SNR at the better ear. For example, in the case of target at -15° and interference at 15°, the initial SNR difference at the two ears is 5.4 dB. Although our model results in only a SNR difference of 0.7 dB between the two ears after segregation, the reconstructed signal corresponding to the better (left) ear contains more target energy, thus yielding better listening quality. Therefore, all the following evaluations are performed at the better ear.

For the two-source case, the model is systematically evaluated at the better ear for various combinations of azimuth angles. We compare the SNR gain obtained by our model against that obtained using the ideal binary mask. For the test corpus of Table II, Fig. 11 shows the results for a spatial separation of 5° and target at azimuth 0°, 40° and 80°. Results are similar across mixtures in the same noise category; hence, we present the averaged result for each category. Very good results are obtained when the target is close to the median plane for an azimuth

separation as small as 5°. Performance degrades when the target source is moved to the side of the head and this is a direct consequence of poorer resolution of the binaural cues at higher azimuth angles. When comparing with the SNR of the initial mixture, there is an average-SNR gain of 13.76 dB for the target in the median plane, and it reduces to 5.04 dB with target at 80°. When the spatial separation increases, excellent results are obtained across all spatial configurations. Figure 12 shows results for target at 0°, 30° and 60° and interference at 30° to the right of target. Similar results are obtained for other spatial configurations. The above performance profiles are in qualitative agreement with human experimental data (Blauert, 1997). Figure 13 shows that the system performs equally well on the Cooke corpus. Fig. 13A gives the results for a 5° azimuth separation and the average improvement is 13.73 dB. Similarly, Fig. 13B gives the results for a 30° separation.

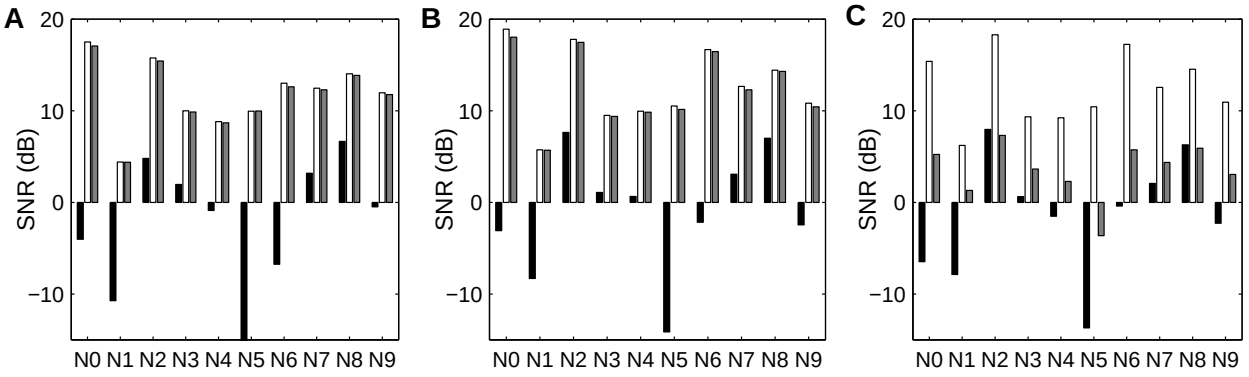

Figure 11. Systematic results for two-source configuration with 5° azimuth separation. Black bars correspond to the SNR of the initial mixture, white bars indicate the SNR obtained using ideal binary mask, and gray bars show the SNR from our model. Results are obtained for speech mixed with ten types intrusions (N0: pure tone; N1: white noise; N2: noise burst; N3: 'cocktail party'; N4: rock music; N5: siren; N6: trill telephone; N7: female speech; N8: male speech; N9: female speech) for different spatial configurations. **A:** Target at 0°, interference at 5°. **B:** Target at 40°, interference at 45°. **C:** Target at 80°, interference at 85°.

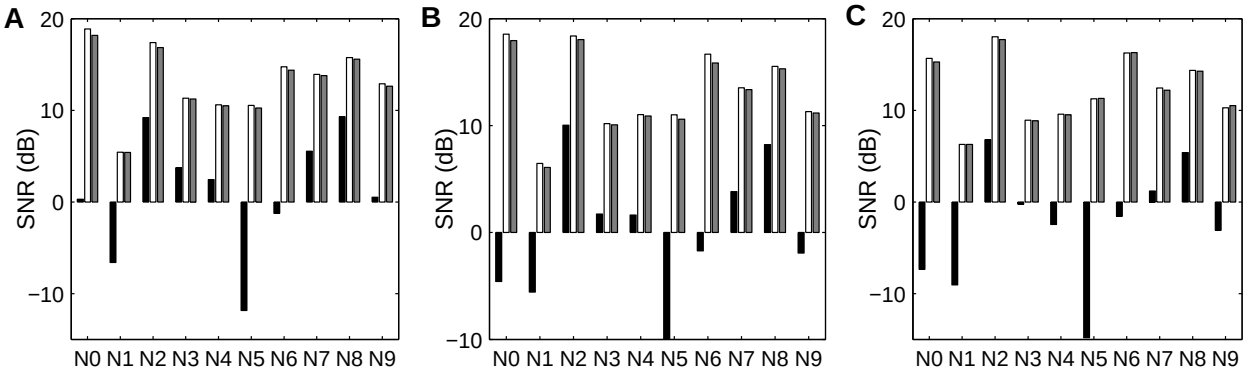

Figure 12. Systematic results for two-source configuration with 30° azimuth separation. Black bars correspond to SNR of the initial mixture, white bars to the SNR obtained using an ideal binary mask, and gray bars to the SNR from our model. **A:** Target at 0°, interference at 30°. **B:** Target at 30°, interference at 60°. **C:** Target at 60°, interference at 90°.

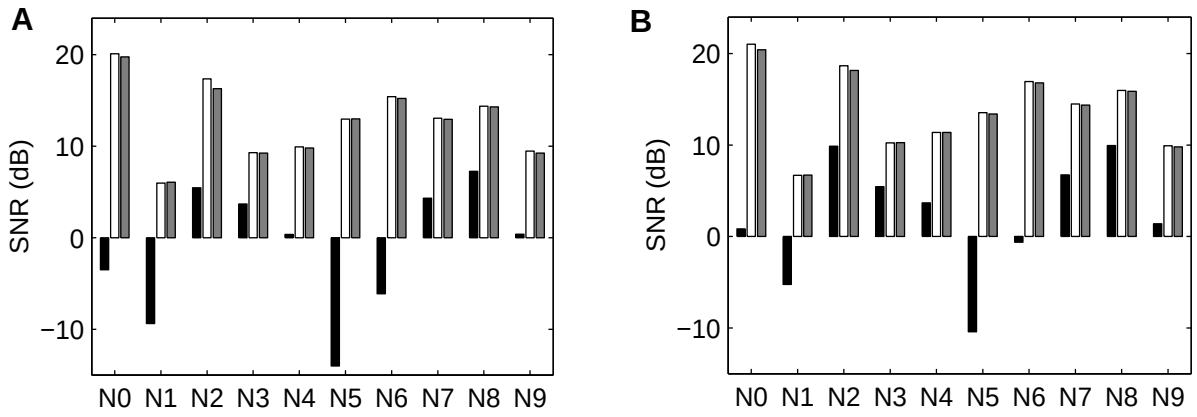

Figure 13. Systematic results for two-source configuration using the Cooke corpus as the test corpus. Black bars correspond to SNR of the initial mixture, white bars to the SNR obtained using an ideal binary mask, and gray bars to the SNR from our model. **A:** Target at 0°, interference at 5°**. B:** Target at 0°, interference at 30°.

Our approach, like other location-based methods using cross-correlation, can be extended to cases with more than two sources. With given locations, our model performs target segregation in a similar manner. Figure 14 illustrates the performance of the model in a three-source scenario with target located in the median plane and two interfering sources at –30° and 30°. Here, the 10 noise intrusions from the Cooke corpus are presented at 30° azimuth and the target is reconstructed based on the right ear mixture (closer to 30°). As previously, results are mean values for the 10 types of noise intrusion. The performance degrades compared to the corresponding two-source situation, from an average SNR of about 12 dB to 4.10 dB. Still, the average SNR gain obtained is approximately 11.31 dB. Informal listening tests suggests that the model filters out the interference effectively.

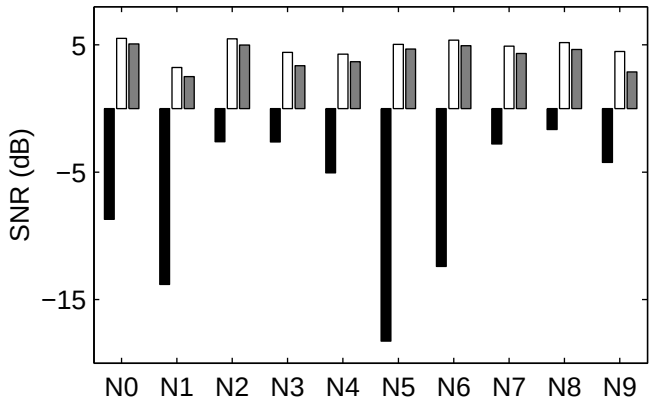

Figure 14. Evaluation for a three-source configuration. The target is in the median plane and intrusions are at -30° and 30°. Black bars correspond to the SNR of the initial mixture, white bars to the SNR obtained using ideal binary mask, and gray bars to the SNR from our model.

In order to draw a quantitative comparison, we have implemented Bodden model (Bodden, 1993), which produces good-quality sound separation using source locations. The method used in

the Bodden's cocktail-party processor is a great deal more complicated than the model presented here. Bodden's system uses a 24-channel filterbank intended to simulate critical bands. His model contains an extended cross-correlation mechanism based on contralateral inhibition that incorporates ITD in the low frequency range and IED in the high frequency range as well as IID. Additional weights in the cross-correlation method are trained to adapt the system to the actual HRTFs. For a fair comparison, our implementation of the Bodden system uses the same 128 channel gammatone filterbank employed in our system; we also implemented the Bodden model with 24-channel critical bands and the results are not as good. We find that, when two sources are relatively close, the Bodden model is less robust than ours. Our comparison is based on the Cooke corpus and a spatial configuration of target at 0° and intrusion on the right side at 30°, an azimuth separation in the range where his model performs optimally. As displayed in Fig. 15, our model shows a considerable improvement over the Bodden system, producing 3.5 dB average improvement.

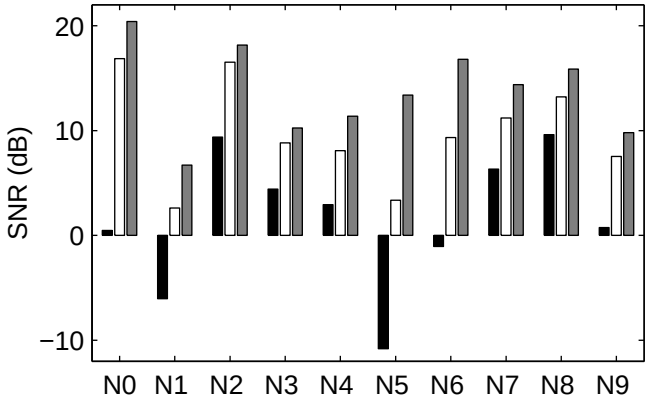

Figure 15. SNR comparison between the Bodden model (white bars) and our model (gray bars) for a two-source configuration: target in the median plane and interference at 30°. The black bars corresponds to the SNR of the original mixture.

DISCUSSION

The human auditory system is capable of adapting to a variety of acoustical situations. A key feature of our model is the introduction of supervised learning for different spatial configurations, and training is conducted independently for different frequency channels. We assume that such training takes place before performing specific segregation tasks, and it would correspond to learning during the development stage. Supervised signals for a spatial configuration of target and intrusion could be supplied in a number of ways, including sound localization, signal estimation from a specific location, and even information extracted from a different modality (e.g. vision). It is worth emphasizing that, unlike a typical supervised learning situation, the training here does not need to capture the specific contents of training signals. As a result the model can be trained equally well using other natural sounds, and estimated distributions generalize in a broad range. In an earlier study (Roman *et al.*, 2002), for example, we employed a different training methodology and a different training corpus, but the system performance was very similar.

While satisfying the demands of an effective computational system, our model is strongly motivated by physiological and psychoacoustical findings regarding the extraction of spatial features (Patterson *et al.*, 1988). The peripheral processing is based on a gammatone filterbank, which has a foundation in physiology and psychoacoustics. Similarly, the cross-correlation mechanism for ITD extraction as well as the across-frequency integration for localization are supported by related physiological findings (Popper and Fay, 1992).

An open question concerns the role of spatial location in perceptual separation of competing sounds. The experiments by Culling and Summerfield (1995), using simulated vowels in which the formants were defined by noise bands, showed that simultaneous grouping across frequencies based on ITD is weak. Later experiments by Darwin and Hukin (1997; 1999) found that ITD plays a weak role in concurrent sound segregation, but a much stronger role in linking acoustic events from a common location over time. The recent experiments of Freyman *et al.* (2001) further showed a sizeable improvement in recognizing target speech in the presence of one or two competing speakers based on perceived spatial separation, which suggests a location-based grouping mechanism. Our computational results demonstrate that computed locations can play an effective role in across-frequency grouping. On the other hand, many monaural cues are also important for sound source segregation (see the Introduction), and how to incorporate both monaural and binaural cues in a comprehensive system remains a challenge.

Our approach uses characteristic clustering of the joint ITD-IID space in order to accurately estimate an ideal binary mask. Related models for estimating target masks have been proposed previously (Glotin *et al.*, 1999; Jourjine *et al.*, 2000). Such models, however, assume input directly from microphone recordings and head-related filtering was not considered. Simulation of human binaural hearing introduces different constraints as well as clues to the problem. First, both ITD and IID should be utilized since IID is more reliable for higher frequencies than ITD. Second, frequency-dependent combinations of ITD and IID arise naturally for a fixed spatial configuration. Consequently, channel-dependent training for each frequency band becomes necessary. Our tests with just ITD (as in Glotin *et al.*) or channel-independent classification (as in Jourjine *et al.*) yield considerably inferior performance.

As illustrated in Fig. 14, the proposed model can be used to extract target speech from an acoustic mixture that contains more than one intrusion. Although segregation results are expected to drop as the number of sources increases, this property of our model differs from blind source separation using independent component analysis (Hyväarinen *et al.*, 2001) or spatial filtering using sensor arrays (Krim and Viberg, 1996); such techniques have strong requirements on the number of sensors that increases as the number of sources increases in an auditory scene. A main reason for this difference is that auditory considerations play a large role in our model design.

In terms of limitations, our model currently does not address room reverberation or moving sound sources. The localization of multiple moving sources in reverberant conditions with just two sensors is a complex topic. Some tracking mechanism based on measurements of binaural cues across frequency channels, combined with channel selection to discard unreliable T-F units, could be employed to estimate the locations of active sources. For voiced sources, periodicity may provide a measure for the reliability of T-F units (see Wu *et al.*, 2002). Other auditory mechanisms, such as the precedence effect and forward/backward masking, could also provide important cues to cope with reverberation. Our model also does not address how to define a target in a multi-source situation, and to

address this issue would inevitably require some high-level processes such as attention and task specification. We plan to investigate these and other related issues in future work.

To conclude, we have proposed a model for speech segregation based on spatial location. We have observed systematic deviations of the ITD and IID cues with respect to the relative strength between target and acoustic interference, and configuration-specific clustering in the joint ITD-IID feature space. Consequently, learning of binaural patterns can be employed for individual frequency channels and different spatial configurations. Finally, the system estimates a binary mask in order to eliminate acoustic energy in time-frequency units where interference is stronger than target. Our model has been systematically evaluated, and it achieves substantial improvement over an existing computational system.

ACKNOWLEDGMENTS

This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-0027). A preliminary version of this work is included in the Proceedings of 2002 ICASSP.

References

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

Bodden, M. (**1993**). "Modeling human sound-source localization and the cocktail-party-effect," *Acta Acoustica*, vol. 1, pp. 43-55.

Bodden, M. (**1996**). "Auditory demonstrations of a cocktail party processor," *Acustica*, vol. 82, pp. 356-357.

Bregman, A. S. (**1990**). *Auditory Scene Analysis*, Cambridge, MA: MIT press.

Bronkhorst, A. (**2000**). "The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions," *Acustica*, vol. 86, pp. 117-128.

Brown, G. J., and Cooke, M. P. (**1994**). "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297-336.

Burkhard, M. D., and Sachs, R. M. (**1975**). "Anthropometric manikin for acoustic research," *J. Acoust. Soc. Am.*, vol. 58, pp. 214-222.

Colburn, H. S. (**1973**). "Theory of binaural interaction based on auditory nerve data. I. General strategy and preliminary results on interaural discrimination," *J. Acoust. Soc. Am.*, vol. 54, pp. 1458-1470.

Cooke, M. P. (**1993**), *Modeling Auditory Processing and Organization*, Cambridge, U.K.: Cambridge University Press.

Cooke, M. P., Green, P., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267-285.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," *J. Acoust. Soc. Am.*, vol. 98, pp. 785-797.

Darwin, C. J., and Hukin, R. W. (**1997**). "Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity," *J. Acoust. Soc. Am.*, vol. 102, pp. 2316-2324.

Darwin, C. J., and Hukin, R. W. (**1999**). "Auditory objects of attention: The role of interaural time differences," *J. Exp. Psychol.:Human Percept. Perform.*, vol. 25, pp.617-629.

Duda, R. O., Peter, E. H., and Stork, D. G. (**2001**), *Pattern Classification*, 2nd Edition, New York, Wiley.

Durlach, N. I. (**1972**). "Binaural signal detection: equalization and cancellation theory," *Foundations of Modern Auditory Theory*, vol.II. , ed. J.V. Tobias, Academic Press.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," *J. Acoust. Soc. Am.*, vol. 109, pp. 2112-2122.

Gaik, W. (**1993**). "Combined evaluation of interaural time and intensity differences: psychoacoustic results and computer modeling," *J. Acoust. Soc. Am.*, vol. 94, pp. 98-110.

Gardner, W. G., and Martin, K. D. (**1994**). "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab Perceptual Computing Technical Report #280.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (**1993**). "Darpa timit acoustic-phonetic continuous speech corpus," Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 44, pp. 99-122.

Glotin, H., Berthommier, F., and Tessier, E. (**1999**). "A CASA-Labelling model using the localisation cue for, robust cocktail-party speech recognition," *Proc. Eurospeech*, vol. 5, pp. 2351 2354.

Hyväarinen, A., Karhunen, J., and Oja, E. (**2001**). *Independent component analysis.* New York: Wiley.

Jeffress, L. A. (**1948**). "A place theory of sound localization," *J. Comp. Physiol. Psychol*., vol. 41, pp. 35-39.

Jourjine, A., Rickard, S., and Yilmaz, O. (**2000**). "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," *Proc. ICASSP*, vol. 5, pp. 2985-2988.

Krim, H., and Viberg, M. (**1996**). "Two decades of array signal processing research: The parametric approach," *IEEE Sig. Proc. Mag.*, vol. 13, pp. 67-94.

Lindemann, W. (**1986**). "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.

Liu, C., Wheeler, B. C., O'Brien, W. D., Jr., Lansing, C. R., Bilger, R. C., Jones, D. L., and Feng, A. S. (**2001**). "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," *J. Acoust. Soc. Am.*, vol. 110, pp. 3218-3230.

Lyon, R. F. (**1983**). "A computational model of binaural localization and separation," *Proc. of IEEE ICASSP*, pp. 1148-1151.

MacPherson, E. A. (**1991**). "A computer model of binaural localization for stereo imaging measurement," *J. Audio Eng. Soc.*, vol. 39, pp. 604-622.

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

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for prediction of thresholds, loudness and partial loudness," *J. Audio Eng. Soc.*, vol. 45, pp 224-240.

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

Popper, A. N., and Fay, R. R. (**1992**). Ed. *The mammalian auditory pathway: Neurophysiology.* New York: Springer-Verlag.

Roman, N., Wang, D. L., and Brown, G. J. (**2002**). "Location-based sound segregation," *Proc. ICASSP*, vol. 1, pp. 1013-1016.

Sain, S. R., Baggerly, K. A., and Scott, D. W. (**1994**). "Cross-Validation of Multivariate Densities," *J. Am. Stat. Assoc.*, vol. 89, pp. 807-817.

Silverman, B. W. (**1986**). *Density estimation for statistics and data analysis*, New York: Chapman and Hall.

Shackleton, T. M., Meddis, R., and Hewitt, M. J. (**1992**). "Across frequency integration in a model of lateralization," *J. Acoust. Soc. Am.*, vol. 91, pp. 2276-2279.

Slatky, H. (**1993**). "Algorithmen zur richtungsselektiven Verarbeitung von Schallsignalen eines binauralen Cocktail-Party-Prozessors", PhD thesis, Ruhr-Universität Bochum.

Stern, R. M., and Trahiotis, C. (**1995**). "Models of binaural interaction," *Hearing*, ed. B.C.J. Moore, Academic Press.

Wang, D. L., and Brown, G. J. (**1999**). "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Net.*, vol. 10, pp. 684-697.

Wu, M., Wang, D. L., and Brown, G. J. (**2002**). "A multipitch tracking algorithm for noisy speech," *Proc. ICASSP*, vol.1, pp. 369-372.