Speech segregation based on sound localization

Nicoleta Roman[a)]
Department of Computer and Information Science, The Ohio State University, Columbus, Ohio 43210

DeLiang Wang[b)]
Department of Computer and Information Science and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

Guy J. Brown[c)]
Department of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom

(Received 25 June 2002; accepted for publication 18 July 2003)

At a cocktail party, one 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, supervised learning approach to speech segregation, in which a target speech signal is separated from interfering sounds using spatial localization cues: interaural time differences (ITD) and interaural intensity differences (IID). Motivated by the auditory masking effect, the notion of an ''ideal'' time–frequency binary mask is suggested, which selects the target if it is stronger than the interference in a local time–frequency (T–F) unit. It is observed that within a narrow frequency band, modifications to the relative strength of the target source with respect to the interference trigger systematic changes for estimated ITD and IID. For a given spatial configuration, this interaction produces characteristic clustering in the binaural feature space. Consequently, pattern classification is performed in order to estimate ideal binary masks. A systematic evaluation in terms of signal-to-noise ratio as well as automatic speech recognition performance shows that the resulting system produces masks very close to ideal binary ones. A quantitative comparison shows that the model yields significant improvement in performance over an existing approach. Furthermore, under certain conditions the model produces large speech intelligibility improvements with normal listeners. [DOI: 10.1121/1.1610463]

PACS numbers: 43.72.Ew, 43.66.Ba, 43.66.Qp [DOS]

I. 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, localization (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 originates 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 for both localization and recognition tasks. Headphone 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 deg. On the other hand, the smallest detectable change in IID by the human auditory system is about 0.5 to 1 dB at all frequencies. Resolution deteriorates as the reference ITD gets larger, and the difference limen can be as much as 10 deg when the ITD corresponds to a source 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 cancellation (EC) theory (Durlach, 1972; Breebaart *et al.*, 2001), and auditory-nerve-based models (Colburn, 1977; Stern and Colburn, 1978). 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; Whittkop and Hohmann, 2003). Most cocktail-party-processor designs utilize the following observation: as

[a)]Electronic mail: niki@cis.ohio-state.edu
[b)]Electronic mail: dwang@cis.ohio-state.edu
[c)]Electronic mail: g.brown@dcs.shef.ac.uk

 0001-4966/2003/114(4)/2236/17/$19.00

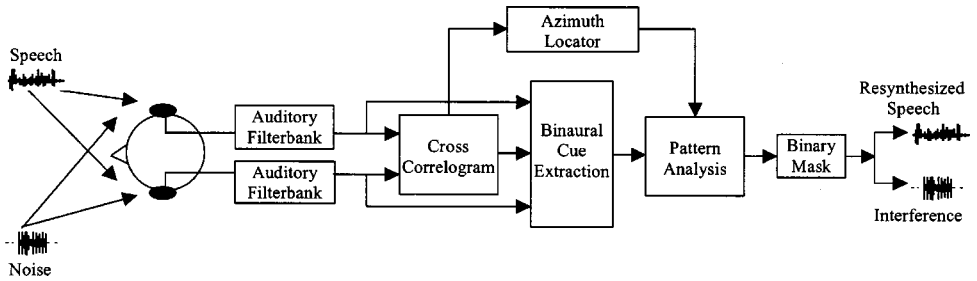

FIG. 1. Schematic diagram of the model. Binaural signals are obtained by convolving input signals with measured head related impulse responses (HRIR) from a KEMAR dummy head. A model of the auditory periphery is employed. Azimuth localization for all the sources is based on a cross-correlation mechanism. 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 relative strength of the interference with respect to the target increases, certain attributes of the auditory event including location and spatial 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 desired cross-correlation patterns to observed ones. 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). Such a model can be applied to, among other things, enhancing speech recognition in noisy environments and improving binaural hearing aid design. 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. Here, we introduce the notion of an ideal binary mask, which is defined *a priori* as follows. An element in the mask is 1 if the corresponding T–F unit contains target energy that is stronger than interference energy, and 0 otherwise. When target and intrusion are available before mixing, as is the case during our evaluation, ideal binary masks can be readily constructed. We call such a mask *ideal* because its construction requires the knowledge of individual sound sources before mixing. In addition, from a theoretical ASA perspective, an ideal binary mask gives a performance ceiling for all binary masks. Note that an ideal mask remains well-defined for situations when more than one target needs to be segregated. The ideal mask notion is motivated by the human auditory masking phenomenon, in which a stronger signal masks a weaker one in the same critical band (Moore, 1997). Moreover, such masks generate high-quality reconstruction for a variety of signals, and have been recently shown to provide a highly effective front end for robust speech recognition (Cooke *et al.*, 2001). Furthermore, as will be shown later, deviations from ideal binary masks lead to gradual degradation in speech recognition performance. Hence, our model aims to estimate an ideal binary mask using information about the spatial configuration of sound sources.

Statistics for the relationship between the relative strength of sources and the pattern of 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 of target and interference and estimated 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 an ideal mask. An objective evaluation of the system with both SNR (signal-to-noise ratio) and ASR (automatic speech recognition) measures shows that the results of our system are comparable with those obtained using ideal binary masks. In addition, a speech intelligibility evaluation using normal listeners shows a large improvement under certain conditions.

The rest of the paper is organized as follows: the next section contains an overview of the model. Section III describes the peripheral auditory model. Section IV describes the azimuth localization algorithm. Section V is mainly devoted to the ideal binary mask estimation, which constitutes the core of the model. Section VI presents the evaluation results of the system and a quantitative comparison with the Bodden model. In the last section we give further discussions and future directions.

II. 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 an ideal binary mask; and (4) reconstruction of the target signal. Figure 1 illustrates the model architecture for the case of two sound sources.

The input to our model is a mixture of two or more signals at different, but fixed, locations: target speech and acoustic interference. Measurements of head-related transfer functions (HRTF) are a standard method for realistic binaural synthesis. We utilize here a catalog of HRTF measurements collected by Gardner and Martin (1994) from a KEMAR dummy head under anechoic conditions. The measurements consist of left/right KEMAR responses from a distance of 1.4 m in the horizontal plane, resulting in 128 point impulse

responses at a sampling rate of 44.1 kHz. Binaural signals are obtained by filtering monaural signals with HRTFs 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.

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, which performs half-wave rectification and square-root compression. The output of the model gives a firing rate representation of auditory-nerve activity.

Simulated 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. Our localization stage uses only ITD information. Consequently, the system cannot tell front from back. We restrict our model to the half-horizontal plane with azimuth in the range $[-90°,90°]$. 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 supervised 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 an ideal binary mask. The time–frequency resolution for the current implementation is 20-ms time frames with a 10-ms frame shift (see, e.g., Wang and Brown, 1999), and 128 frequency channels that cover the range of 80 Hz to 5 kHz.

The last stage of the model is a reconstruction path, which allows the target signal to be recovered from the acoustic mixture by nullifying the T–F units dominated by interference. The method employed here is the same in principle as that described by Weintraub (1986) (see also Brown and Cooke, 1994). The target signal is reconstructed from the output of the gammatone filterbank. To remove across-channel phase differences, the output of a filter is time reversed, passed through the gammatone filter, and time reversed again. Furthermore, the output for each filter is divided in 20-ms sections with 10-ms overlap that correspond to T–F units in the binary mask, and windowed with a raised cosine. Binary weights estimated in the previous stage are then applied to each section to remove the interference. This method achieves high-quality reconstruction (Weintraub, 1986; Brown and Cooke, 1994; Wang and Brown, 1999).

III. AUDITORY PERIPHERY

It is widely acknowledged that cochlear filtering can be modeled by a bank of bandpass filters. 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:

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

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 of auditory filter bandwidth. 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):

$$\mathrm{ERB}(f_i) = 24.7(4.37 f_i/1000 + 1), \tag{2}$$

$$b_i = 1.019\,\mathrm{ERB}(f_i). \tag{3}$$

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 are provided by Moore *et al.* (1997). We include this middle-ear processing for the purpose of physiological plausibility. 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 rectified signal.

Psychophysical models for sound localization generally employ envelopes of the responses in the high-frequency range. This is supported by discrimination experiments using transposed stimuli, suggesting similar sensitivity to ITD for both low- and high-frequency ranges (Bernstein and Trahiotis, 2001). Therefore, we additionally extract the envelopes using the Hilbert transform for channels with center frequencies above 1.5 kHz. Note that the envelope is not actually used in our current implementation; rather, it is used in Sec. V as part of a comparison of the effectiveness of different interaural cues.

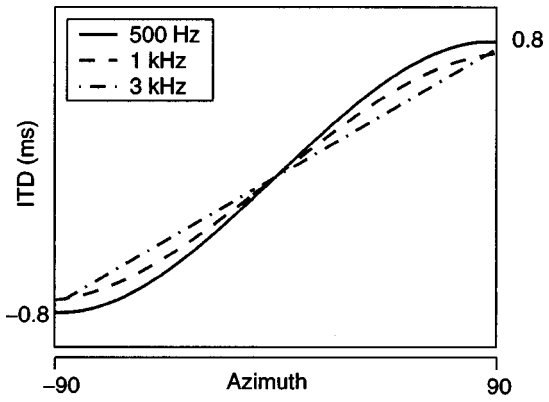

FIG. 2. Functions relating azimuth to ITD for three auditory channels with center frequencies of 500 Hz, 1 kHz, and 3 kHz.

IV. AZIMUTH LOCALIZATION

Current models of azimuth localization almost invariably employ cross correlation, which is functionally equivalent to the coincidence detection mechanism proposed by Jeffress (1948). Cross correlation provides excellent time delay estimation for broadband stimuli, and for narrow band stimuli in the low-frequency range. However, for high-frequency narrow band signals it produces multiple ambiguous peaks. Here, we use the normalized cross correlation computed at lags equally distributed from -1 to 1 ms ($-44 < \tau < 44$) using a rectangular integration window of 20 ms (corresponding to $K = 880$ samples). 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:

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

where l_i, r_i refer to the left and right auditory periphery output of the ith channel, and $\overline{l_i}$, $\overline{r_i}$ 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 the 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. The mappings are obtained based on the cross-correlation output in response to white noise presented systematically at locations in the azimuth range $[-90°,90°]$. Figure 2 shows three ITD-azimuth mappings, for channels with center frequencies of 500 Hz, 1 kHz, and 3 kHz. The functions are monotonic, being sigmoidal at low frequencies where diffraction effects are greater and increasingly linear at high frequencies.

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. That is, each local peak generates an impulse of the same height and then the obtained impulse train is convolved with a Gaussian. Here, the width is linear with respect to the center frequency of the channel. This technique sharpens the cross-correlogram, an effect similar to a lateral inhibition mechanism (Arbib, 2003).

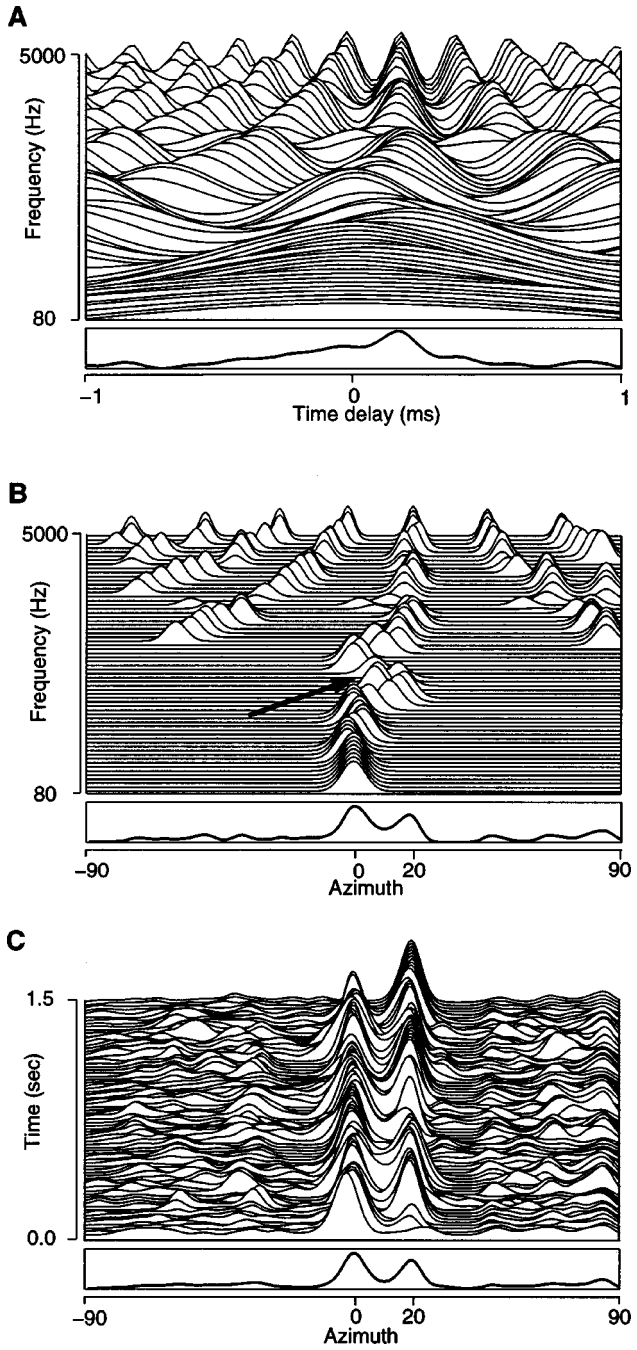

FIG. 3. Azimuth localization for a mixture of male utterance at 0° and female utterance at 20°. The bottom plot in each panel shows a summation across all rows. (A) Cross-correlation functions for 128 frequency channels in the range 80 Hz–5 kHz at time frame 40 (i.e., 400 ms after the start of the stimulus). 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 s, shown every 20 ms.

The cross-correlation method provides inconsistent results when two acoustic sources are present. Figure 3 shows

the cross-correlation functions [Fig. 3(A)] and the skeleton cross-correlogram [Fig. 3(B)] for a mixture of male speech presented at 0° and female speech presented at 20°. Here, the width of the Gaussians in the skeleton cross-correlogram ranges from 4° at the low-frequency end to 2° at the high-frequency end. For frequency channels where one source is much stronger, 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. Summing across frequencies produces sharper peaks on the skeleton cross-correlogram; in the case of Fig. 3, the skeleton cross-correlogram gives a good estimate of source locations, whereas the conventional cross-correlogram does not [compare the bottom plots in Fig. 3(A) and Fig. 3(B)]. In Fig. 3(C) we display the pooled cross-correlogram for a signal of duration 150 frames (i.e., 1.5 s). 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. 3(C). This represents our method for azimuth localization.

V. IDEAL MASK ESTIMATION

The objective of this stage of the model is to develop an efficient mechanism for estimating an ideal binary mask, which selects the T–F units where the estimated signal energy is greater than the noise energy (i.e., greater than 0-dB SNR). Note that different SNR criteria are possible for defining an ideal binary mask (see Cooke *et al.*, 2001). In the absence of evidence for a better SNR measure, we choose the 0-dB criterion for simplicity. 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 (see Slatky, 1993, for an extensive study of binaural cues with sinusoidal signals). Although in real-world scenarios the conditions of this simplified model are generally not fulfilled, our experimental results show that a similar trend holds for a variety of natural signals when analyzed in narrow frequency bands. This analysis also serves to motivate the introduction of our proposed algorithm for the general case in subsection B.

A. Pure tones

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

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

where A_i is the amplitude, ω_i is the frequency, d_i corresponds to the interaural time delay (equivalent to the phase difference between left and right HRTFs at frequency ω_i), and $H_i^r(\omega_i)$ and $H_i^l(\omega_i)$ represent, respectively, the right and left HRTF, for the ith 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 only in a narrow band 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 trade-off 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.

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

In this case, we have

$$\begin{aligned} c(\tau)&=\frac{A_1^2}{2}\cos(\omega(\tau-d_1))+\frac{A_2^2}{2}\cos(\omega(\tau-d_2))\\ &\quad+A_1A_2\cos\left(\omega\left(\tau-\frac{d_1+d_2}{2}\right)\right)\\ &\quad\cdot\cos\left(\Delta\varphi+\omega\frac{d_2-d_1}{2}\right). \end{aligned} \tag{8}$$

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 ωd_1, ωd_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 $R=A_2/(A_1+A_2) \in [0,1]$. Hence, we derive the solution for $\tau_{\max}$ as follows:

$$\tau_{\max}=\frac{d_1+d_2}{2} + \frac{1}{\omega}\left(\arctan\left[\frac{(A_2^2-A_1^2)\sin\beta}{(A_1^2+A_2^2)\cos\beta+2A_1A_2\cos(\Delta\varphi+\beta)} \right] + k\pi \right), \tag{10}$$

where $\beta=\omega[(d_2-d_1)/2] \in [0,\pi]$ and k is an integer. The relation obtained in (9) uniquely determines $k \in \{0,\pm 1\}$ for the 2π interval considered. More specifically, $\beta \leqslant \pi/2 \Rightarrow k=0$ and $\beta > \pi/2 \Rightarrow k=1$ when $A_1 < A_2$, and $k=-1$ when $A_1 > A_2$. Furthermore, simulations and derivations show that a good approximation for the mean value $\overline{\tau}_{\max}$ when $\Delta\varphi$ varies uniformly in the range $[-\pi,\pi]$ is given by

$$\overline{\tau}_{\max} \approx \begin{cases} d_1, & R<0.5 \\ \dfrac{d_1+d_2}{2}, & R=0.5 \\ d_2, & R>0.5. \end{cases} \tag{11}$$

2. 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

$$c(\tau)=\frac{A_1^2}{2}\cos(\omega_1(\tau-d_1))+\frac{A_2^2}{2}\cos(\omega_2(\tau-d_2)). \tag{12}$$

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: $\tau_{\max}=(A_1^2\omega_1^2 d_1+A_2^2\omega_2^2 d_2)/(A_1^2\omega_1^2+A_2^2\omega_2^2)$. Note that this is a monotonic function with respect to the relative strength R.

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

$$\overline{\tau}_{\max}=\frac{d_1+d_2}{2}+\frac{1}{\omega}\left(\arctan\left[\frac{(A_2^2-A_1^2)}{(A_1^2+A_2^2)}\tan\beta \right] + k\pi \right), \quad k \in \{0,\pm 1\}, \tag{13}$$

which is the solution for the maximum position in (12) when $\omega_1=\omega_2$. This function is monotonically increasing with respect to R when $\beta < \pi/2$ and decreasing when $\beta > \pi/2$. Figure 4 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 complex and not amenable to the analysis used here. 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:

$$\mathrm{IID}=10\log_{10}\frac{A_1^2|H_1^r(\omega_1)|^2+A_2^2|H_2^r(\omega_2)|^2}{A_1^2|H_1^l(\omega_1)|^2+A_2^2|H_2^l(\omega_2)|^2}, \tag{14}$$

where the power of a signal $u(t)$ is $\lim_{T\to\infty} 1/2T \int_{-T}^{T} u^2(t)dt$. 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 $[\omega-\Delta\omega,\omega+\Delta\omega]$. Figures 5(A) and (B) 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 to 200 Hz. In the figure, M_1 is the ITD mean as derived in (11) and it approximates well the case $\Delta\omega=0$. M_2 is the ITD mean derived in (13) for the more general case $\Delta\omega \neq 0$. Similarly, Figs. 5(C) and (D) show results for IID when $\omega=2.5$ kHz and five $\Delta\omega$ values in the range of 0 to 400 Hz. Here, M is the IID mean as derived in (14). 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

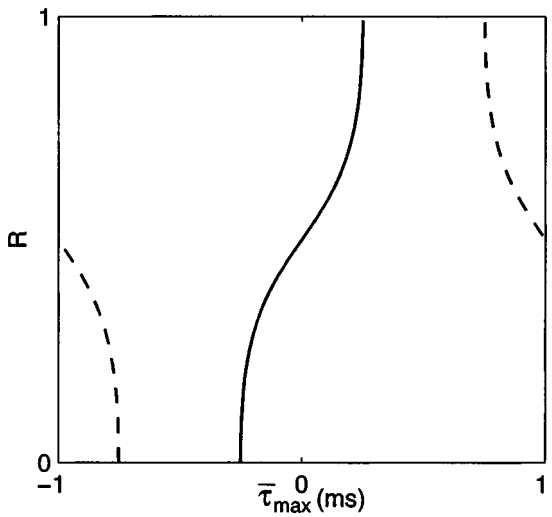

FIG. 4. Theoretical approximation for the mean ITD, $\overline{\tau}_{\max}$, for two pure tones randomly distributed in a narrow band centered at 500 Hz. The *y* axis corresponds to the relative strength *R*. Two cases are shown: $\beta=\pi/4$ (solid line) and $\beta=3\pi/4$ (dashed line).

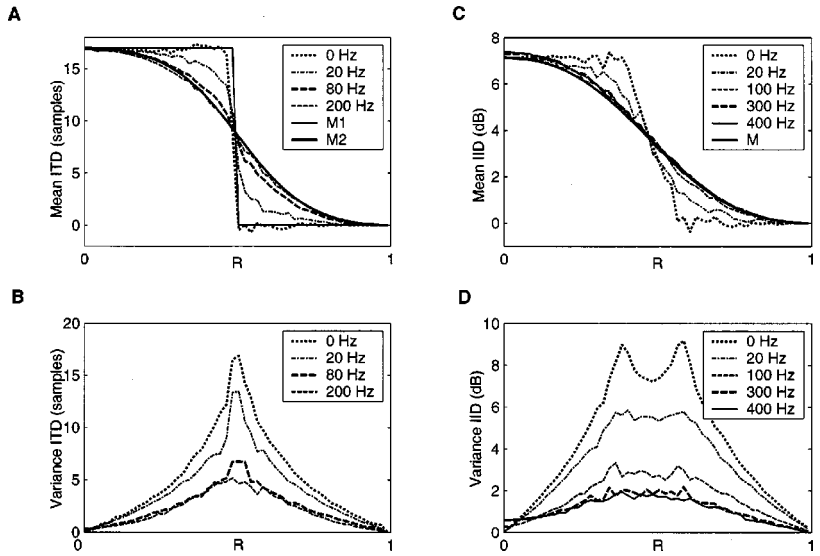

FIG. 5. The influence of filter bandwidth on the mean and variance of ITD and IID with respect to the relative strength R. The data are from simulations of two pure tones uniformly distributed in a narrow band. One tone is at 0° and the another is at 30°. The sampling frequency is 44.1 kHz. (A) Mean ITD as a function of R for 500-Hz center frequency and four bandwidths between 0 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 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).

comparison of the above theoretical derivations with the real data presented in the next subsection shows that the match is very close.

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. Binaural signals are obtained by convolving with KEMAR HRTFs as explained in Sec. II. We employ a training corpus consisting of ten speech signals from the TIMIT database (Garofolo *et al.*, 1993): five male utterances and five 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 the target and the rest interference. In the three-source case, we have S0–S3 as target signals and the two interfering sets are S4–S6 and S7–S9.

Estimates for ITD, IID, and R are extracted independently for all frequency channels. 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 Sec. IV. 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

$$\mathrm{IID}_i = 20\log_{10}\left(\sum_t r_i^2(t) \Big/ \sum_t l_i^2(t)\right), \tag{15}$$

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 with higher SNR (see, e.g., Shinn-Cunningham *et al.*, 2001)

$$R_i = \sqrt{\sum_t s_i^2(t)} \Big/ \left(\sqrt{\sum_t s_i^2(t)} + \sqrt{\sum_t n_i^2(t)}\right), \tag{16}$$

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).

Figure 6 shows empirical results obtained for a two-source configuration: target source in the median plane and interference at 30°. The scatter plot in Fig. 6(A) shows samples of ITD_i and R_i obtained for the channel with a center frequency of 500 Hz (about 7000 samples in total). In

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"

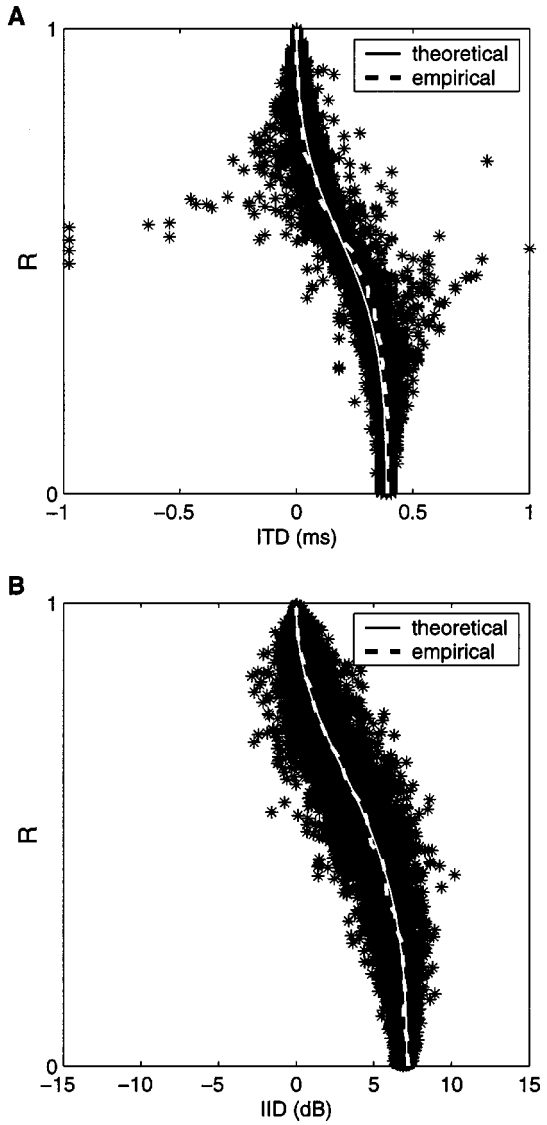

FIG. 6. Relationship between ITD/IID and the relative strength R for a two-source configuration: target in the median plane and interference on the right side at 30°. (A) The scatter plot shows ITD and R estimates from the training corpus for a channel with center frequency of 500 Hz. The solid curve 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.

addition, we display the empirical mean of the samples and the theoretical one derived in (13). Similarly, Fig. 6(B) shows the results that describe the variation of IID_i with R_i for a channel with a center frequency of 2.5 kHz and compares the empirical mean with the one derived in (14). Note that R_i incorporates the HRTF responses at the better ear. Therefore, the R axis for the theoretical mean is converted accordingly. Figure 6 exhibits a systematic shift of the estimated 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-distracter scenarios. As an example, Fig. 7 displays smoothed histograms that show the relationship between R_i and both ITD_i [Fig. 7(A)] and IID_i [Fig. 7(B)] 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. 7 suggest a low probability for this case. Figure 7 instead shows a clustering phenomenon, suggesting that in most cases only one source dominates a T–F unit.

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. 7(C). There exists a trade-off 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. 8 later). 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

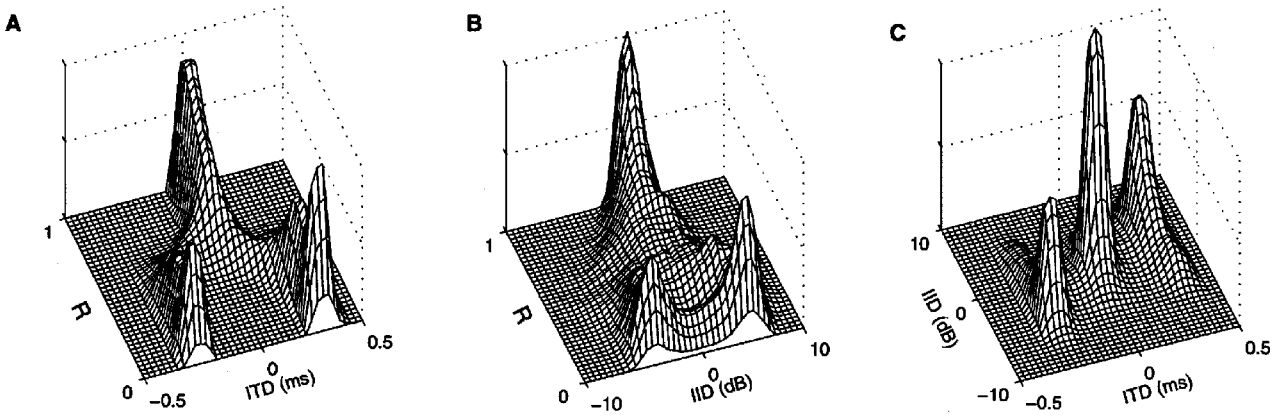

FIG. 7. Relationship between ITD/IID and the relative strength R for a three-source configuration: target source in the median plane and interference at −30° and 30°. Statistics are obtained from the training corpus for a channel with center frequency close to 1.5 kHz. (A) Histogram of ITD and R samples. (B) Histogram of IID and R samples. (C) Clustering in the ITD–IID space.

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 two-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 $R_i \leqslant 0.5$. Based on estimates of the bivariate densities $p(x|H_1)$ and $p(x|H_2)$, the classification is done in accordance with the *maximum a posteriori* (MAP) decision rule: $p(H_1)p(x|H_1) > p(H_2)p(x|H_2)$. There exists 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 as robust due to the following factors: (i) the true number of mixing components is usually unknown, and (ii) the algorithm tends to be sensitive to parameter initialization. Even for the two-source scenario, the method of computing ITD for mid- to high frequencies can result in two-mode distribution for the H_2 hypothesis. In order to completely characterize the distribution of the data, we use the kernel density estimation method independently for all frequency channels.

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

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{n h_1 \ldots h_d} \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right), \tag{17}$$

where $\mathbf{x} = (x_1,\ldots,x_d)$ 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.

One cue not employed in our model is 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. In addition, the occurrence of multiple peaks at high frequencies in the cross-correlation function is much reduced for the IED cue. We have compared the individual performance of the three binaural cues: ITD, IID, and IED, for a one-dimensional classification task based on the kernel density estimation method presented above. An error is made whenever the estimated binary mask value for a T–F unit differs from the corresponding ideal value. Figure 8 shows the error rates with respect to frequency channel using the Cooke corpus (see Sec. VI A) as the test set, where we consider two cases: target source in the median plane and the acoustic interference at 5° [Fig. 8(A)] and 30° [Fig. 8(B)]. 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. 8(B)]. Figure 8 also displays the corresponding error rates for the joint ITD–IID space and the joint IED–IID space, and it shows that the joint ITD–IID space yields the best overall performance across different spatial configurations. As indicated in Fig. 8, we have found no benefit for using IED after incorporating ITD and IID, and hence it is not utilized in our model.

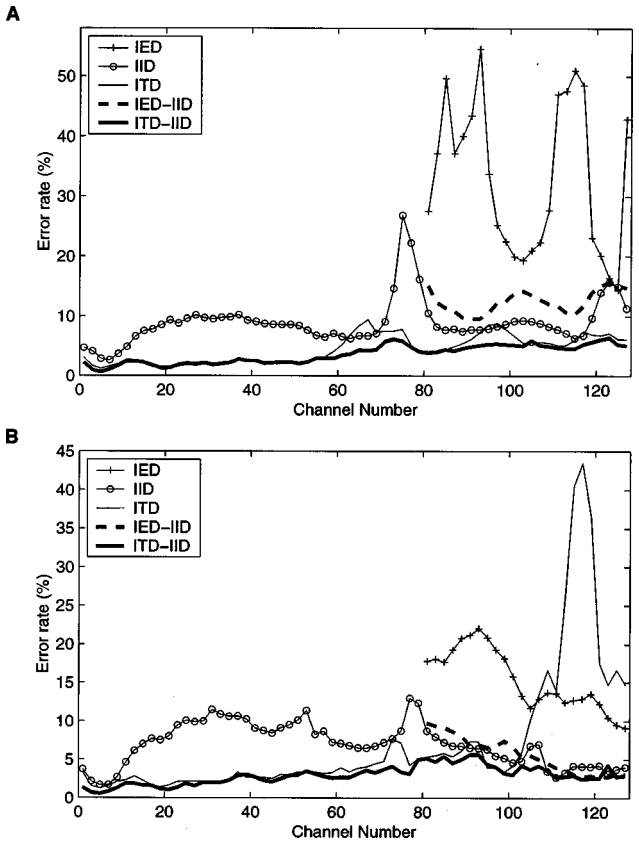

FIG. 8. Discriminability comparison for the three binaural cues, ITD, IID, and IED, the joint IED–IID space, and the joint ITD–IID space. Error rates are displayed as a function of channel number (frequency) for a classification task for two spatial configurations. (A) Target source in the median plane 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.

VI. EVALUATION AND COMPARISON

A binary mask produced by the model described in the last section approximates very well the corresponding ideal binary mask, which is obtained by comparing the energies of

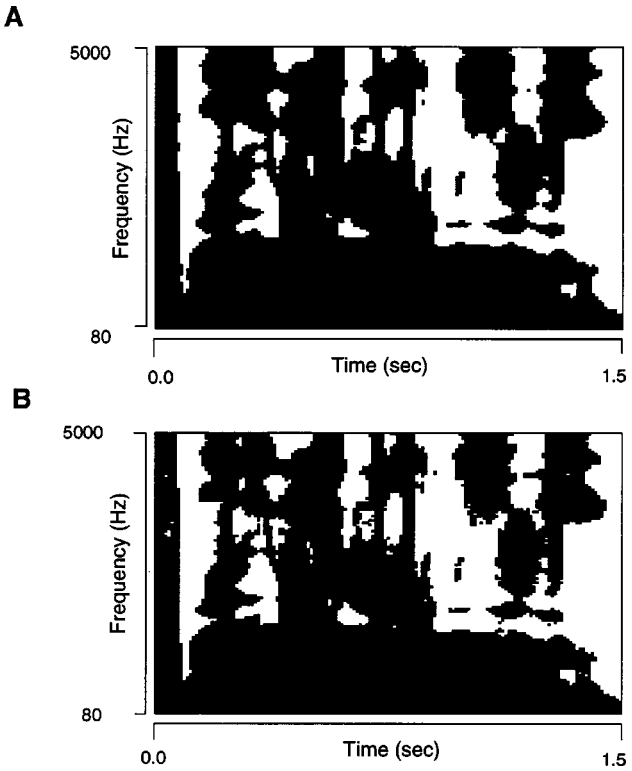

FIG. 9. A comparison between an ideal binary mask (A) and the binary mask resulting from our model (B) for a mixture of male utterance in the median plane (target) and female utterance on the right side at 30° (interference). The black regions indicate those T–F units dominated by target speech.

the original target and interference before mixing. As an example, Fig. 9 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, with an SNR difference of only 0.19 dB.

The performance of a segregation system can be assessed in different ways, depending on intended applications. To extensively evaluate our model, we use the following three criteria: (1) an SNR measure using the original target as signal; (2) ASR rates using our model as a front end; and (3) human speech intelligibility tests. Results with each criterion are given below.

A. SNR evaluation

To conduct an SNR evaluation, a segregated signal is reconstructed from a binary mask following the method described in Sec. II. To quantitatively assess system performance, we measure in decibels the SNR using the original target speech before mixing as signal

$$\mathrm{SNR}=10\log_{10}\sum_t s_T^2(t)\Big/\sum_t (s_T(t)-s_E(t))^2, \quad (18)$$

where $s_T(t)$ represents the original target signal reconstructed using an all-one mask and $s_E(t)$ the estimated target reconstructed from the binary mask. With a binary mask, a more conventional SNR measure would use the mask to pass through original target and intrusion in order to obtain signal and noise, as done by Wang and Brown (1999). The problem with such a measure is that loss of target energy is not penalized, and as a result a separate measure of retained target energy needs to be given (Wang and Brown, 1999). Equation (18) provides a single measure, and in the case of an all-one mask yields the original SNR. Though the signal part in (18) is higher than that retained by a binary mask, it is offset by the denominator that is also higher than retained noise energy; the denominator penalizes both retained noise by the binary mask and target distortion. Our measure is more stringent than the conventional SNR measure; indeed, our tests show that (18) gives systematically lower SNR values. To minimize the loss of target energy we take advantage of the higher initial SNR at the better ear. As a result, the reconstructed signal corresponding to the better ear contains more target energy. Therefore, all the following evaluations are performed at the better ear.

The system performance is measured on independent test corpora for different spatial configurations. 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.*, 2003). The corpus contains ten voiced speech signals and ten 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 ten normal speech utterances from the TIMIT database (see Table II) as target mixed with the ten intrusions from the Cooke corpus (see Table III). In the case of three sources, we use the Cooke corpus for testing: five speech signals form the target set and the other five form one interference source. The ten intru-

TABLE II. Target signals of the test 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"

TABLE III. Noise signals of the test set.

ID	Utterance
N0	1-kHz tone
N1	Random noise
N2	Noise bursts
N3	"Cocktail" party noise
N4	Rock music
N5	Siren
N6	Telephone trill
N7	"Don't ask me to carry an oily rag like..."
N8	"She had your dark suit in greasy wash..."
N9	"Why were we keen to use human..."

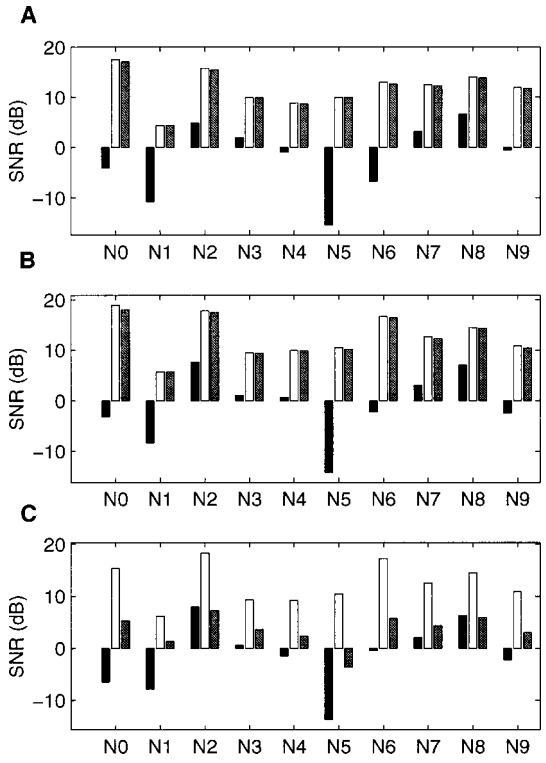

FIG. 10. Systematic results for two-source configuration with 5° azimuth separation. Black bars correspond to the SNR of the initial mixture, white bars indicate the SNR obtained using ideal binary mask, and gray bars show the SNR from our model. Results are obtained for speech mixed with ten types of intrusions (see Table III) 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°.

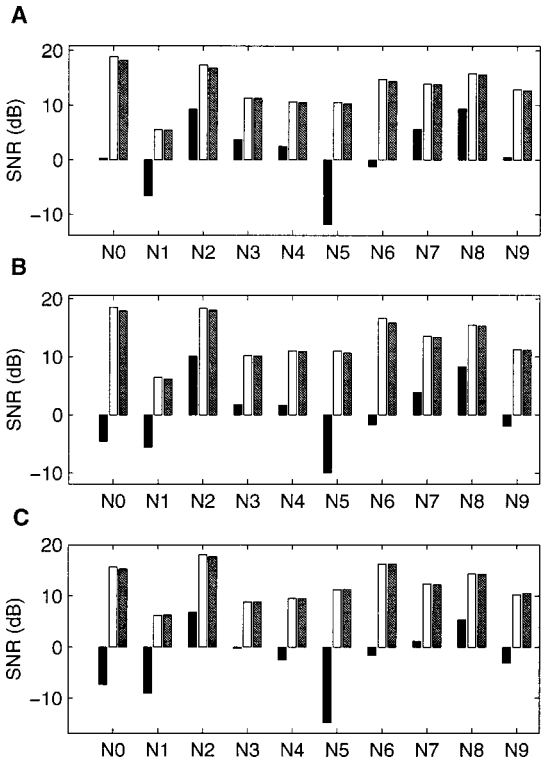

FIG. 11. Systematic results for two-source configuration with 30° azimuth separation. Black bars correspond to SNR of the initial mixture, white bars to the SNR obtained using an ideal binary mask, and gray bars to the SNR from our model. (A) Target at 0°, interference at 30°. (B) Target at 30°, interference at 60°. (C) Target at 60°, interference at 90°.

sions then form the second interference source. Therefore, in this three-source corpus every mixture contains two utterances plus an additional intrusion.

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 an ideal binary mask. For the test corpus of Table II, Fig. 10 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; 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 the target at 80°. When the spatial separation increases, excellent results are obtained across all spatial configurations. Figure 11 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. Figure 12 shows that the system performs equally well on the Cooke corpus. Figure 12(A) gives the results for a 5° azimuth separation and the average improvement is 13.73 dB. Similarly, Fig. 12(B) gives the results for a 30° separation.

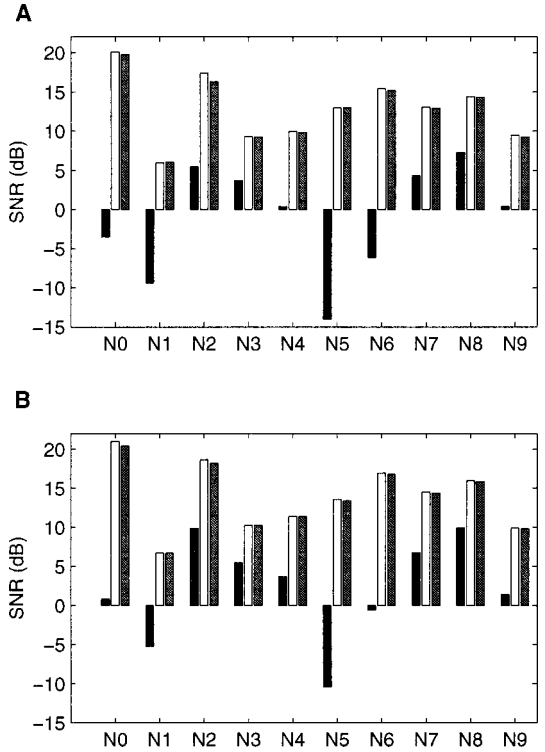

FIG. 12. Systematic results for two-source configuration using the Cooke corpus as the test corpus. Black bars correspond to SNR of the initial mixture, white bars to the SNR obtained using an ideal binary mask, and gray bars to the SNR from our model. (A) Target at 0°, interference at 5°. (B) Target at 0°, interference at 30°.

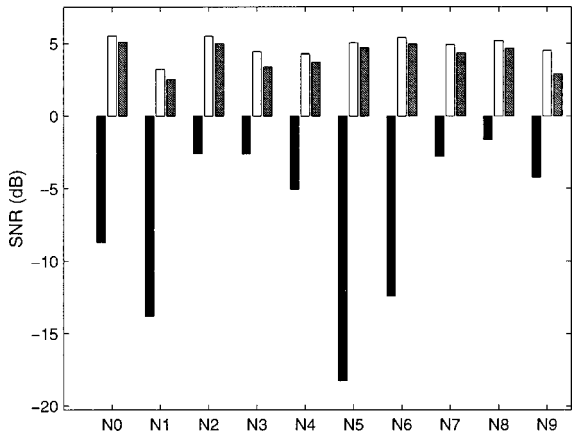

FIG. 13. Evaluation for a three-source configuration. The target is in the median plane and intrusions are at −30° and 30°. Black bars correspond to the SNR of the initial mixture, white bars to the SNR obtained using ideal binary mask, and gray bars to the SNR from our model.

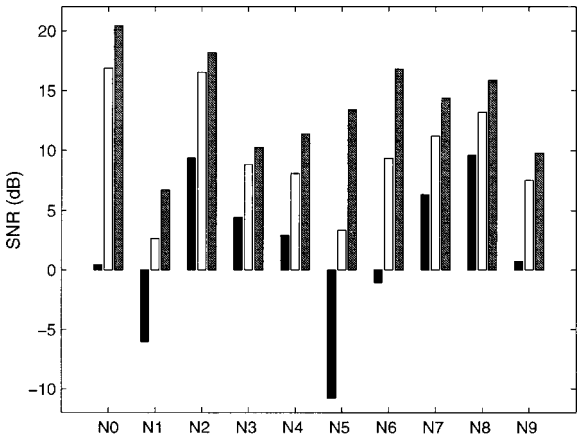

FIG. 14. SNR comparison between the Bodden model (white bars) and our model (gray bars) for a two-source configuration: target in the median plane and interference at 30°. The black bars correspond to the SNR of the original mixture.

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, i.e., estimating an ideal binary mask following the method outlined in Sec. V B. Figure 13 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 ten noise intrusions from the Cooke corpus are presented at 30° azimuth and the target is reconstructed based on the right ear mixture. As previously, results are mean values for the ten types of noise intrusion. The performance degrades compared to the corresponding two-source situation, from an average SNR of about 12 to 4.10 dB. Still, the average SNR gain obtained is approximately 11.31 dB.

In order to draw a quantitative comparison with another binaural processing model, we have implemented the Bodden model (Bodden, 1993), which produces good-quality sound separation using source locations. The localization stage of this model uses an extended cross-correlation mechanism based on contralateral inhibition and it adapts to HRTFs. The separation stage of the model is based on estimation of the weights for a Wiener filter. Specifically, for a given T–F unit the weight is given by the ratio between a desired excitation and an actual one. The actual excitation corresponds to the integration of the cross-correlation pattern across the azimuth axis, and the ideal peak shape is used as a window to derive the desired excitation. The Bodden model differs from ours in several aspects. First, his sound localization stage builds on the previous models of Lindemann (1986) and Gaik (1993), which simulate aspects of the precedence effect for reverberant scenarios, whereas our localization stage is simpler and does not address the precedence effect. Second, his model requires only a target azimuth and no training is necessary as spatial configuration changes. Although these aspects add to the flexibility of his model, the estimation of Wiener filter weights appears less robust than our binary estimation of ideal masks. In addition, our configuration- and channel-specific training utilizes more information provided by localization and makes an optimal use of frequency-dependent ITD and IID cues.

Bodden's system uses a 24-channel filterbank intended to simulate critical bands. For a fair comparison, our implementation of the Bodden system uses the same time–frequency resolution employed in our system with a 128-channel gammatone filterbank; 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. 14, our model shows a considerable improvement over the Bodden system, producing 3.5-dB average improvement. The improvement is especially high for a few cases (e.g., N5 and N6) where our estimated masks result in large SNR improvements over the original mixtures.

B. ASR evaluation

As discussed before, an ideal binary mask is defined *a priori*. Similar *a priori* masks have been shown to produce impressive performance when applied to the automatic recognition of noisy speech using a ''missing data'' approach (Cooke *et al.*, 2001). In this approach, a continuous density hidden Markov model recognizer is modified such that only acoustic features indicated as reliable in the mask are used during decoding. Since our ideal binary masks are generated in a similar way to those used by Cooke *et al.*, we would expect them to be an equally effective front end to missing-data ASR.

Our motivation for ASR evaluation is twofold. First, a practical system must estimate such a mask, and as a result deviations from an ideal mask must be considered. Hence, we want to find how tolerant recognition performance is to deviations from an ideal mask. Second, we want to give a quantitative measure of the potential improvement on ASR performance using our speech segregation model as a front end.

We use the missing-data technique (Cooke *et al.*, 2001) for our ASR evaluation. This technique uses a binary time–frequency mask, where 1's indicate reliable T–F units and 0's unreliable or missing ones. Hence, it works seamlessly with the output from our speech segregation system. We have implemented the missing data algorithm with the same 128-channel gammatone filterbank as described in Sec. III. Feature vectors are obtained using the instantaneous Hilbert envelope at the output of each gammatone filter. More specifically, each feature vector is extracted by smoothing the envelope using an 8-ms first-order filter, sampling at a frame rate of 10 ms and finally log compressing. There are different classification methods for missing-data recognition. Here, we use the bounded marginalization method (Cooke *et al.*, 2001). As in the original study, the task domain is recognition of connected digits, and both training and testing are performed using the male speaker dataset in the TIDigits database (Leonard, 1984).

To study the sensitivity of an ideal mask to estimation error, our first test assesses the correctness score and the accuracy score (correctness minus word insertion errors) when a random deviation from an ideal binary mask is introduced. Here, we use for simplicity a monaural condition as in Cooke *et al.* (2001). Deviations are obtained by randomly flipping the same number of bits from 0's and 1's; the number is measured as percentage of the total number of 1's in an ideal mask. The percentages tested are 0%, 5%, 10%, 20%, and 50%. Since the underlying acoustic energy associated with a T–F unit, or a bit, can vary in a large range, we further measure the energy deviation ratio as the ratio of the energy corresponding to flipped bits and the total energy corresponding to the ideal binary mask. The results for a male target speaker mixed with ''car noise'' (Cooke *et al.*, 2001) are given in Fig. 15, where the abscissa indicates the energy deviation ratio. Three SNR levels for the mixture, i.e., −5, 0, and 5 dB, are tested. Figure 15(A) give the correctness score and Fig. 15(B) the accuracy score. Figure 15 shows that both correctness score and accuracy score decrease gradually and systematically as deviation ratio increases. This suggests that ideal binary masks are robust to estimation error. A comparison between Fig. 15(A) and Fig. 15(B) shows that the accuracy score degrades faster than the correctness score. This suggests that word insertions, which result from noise retention or word boundary blurring, are more sensitive to estimation error than recognition of present words.

The second test directly evaluates binary masks estimated by our system for binaural conditions with two sources and three sources. For all tests, the same male target speaker is located at 0°. Both training and testing of the system are performed on acoustic features from the left ear signal (see Fig. 1). Figure 16(A) and Fig. 16(B) show the correctness and accuracy scores for a two-source condition, where the interference is another male speaker at 30°. The performance of our model is compared against the ideal masks systematically for four SNR levels, i.e., 5, 0, −5, and −10 dB. Also shown in the figure is the baseline performance where the recognition is conducted on unprocessed mixtures from the left ear. Similarly, Fig. 16(C) and Fig. 16(D) show the results for the three-source case with an added female speaker at −30°. The results in Fig. 16 show that an ideal binary mask exhibits only slight and gradual degradation in recognition performance with decreasing SNR and increasing number of sources. In the two-source case, the estimated masks perform equally well as the ideal masks. In the three-source case, the estimated masks do not perform as well, and this is to be expected since we know from Sec. VI A that the quality of ideal mask estimation for three sources is not as good as for two sources. Consistent with the observations from Fig. 15, performance degrades more quickly for the accuracy score than for the correctness score. Observe that large improvements over baseline performance are obtained across all conditions (to a lesser degree for the accuracy score in the three-speaker condition). This shows the strong potential of applying our model to robust speech recognition.

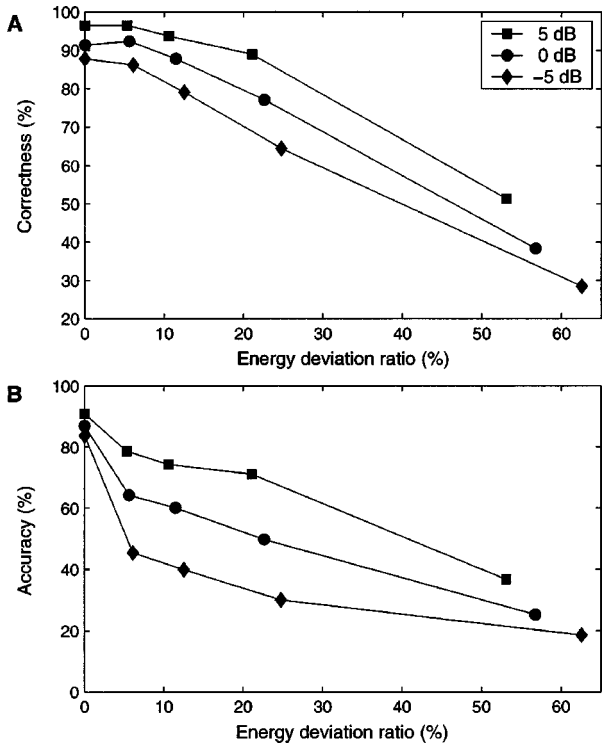

FIG. 15. Degradation of recognition score with deviations from an ideal binary mask evaluated for three SNR values: 5 dB (square), 0 dB (circle), and −5 dB (diamond). (A) Correctness score. (B) Accuracy score.

C. Speech intelligibility evaluation

Finally, we evaluate our model on speech intelligibility with human listeners. Before reporting the results, we should point out that human listeners have a remarkable ability to perform ASA, and their superior ability to recognize speech in the presence of acoustic interference is the very motivation for our model design. Because of this, our tests focus on relatively low SNR conditions; otherwise, scores will be indiscriminately high for both unprocessed mixtures and segregated speech.

We use the Bamford–Kowal–Bench sentence database that contains short semantically predictable sentences (Bench and Bamford, 1979) for intelligibility tests. The score is evaluated as the percentage of keywords correctly identified,

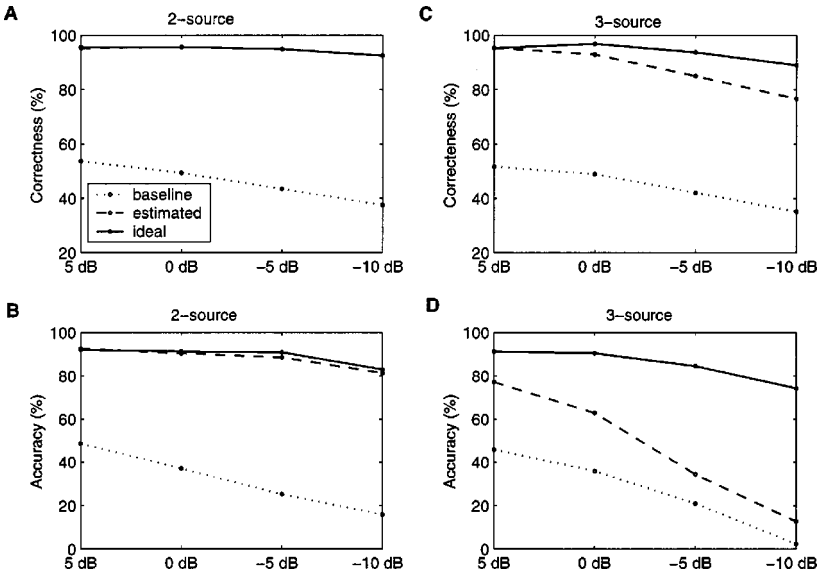

FIG. 16. Recognition performance at different SNR values for original mixture (dotted line), ideal binary mask (solid line), and estimated mask (dashed line). (A) Correctness score for a two-source case. (B) Accuracy score for a two-source case. (C) Correctness score for a three-source case. (D) Accuracy score for a three-source case.

ignoring minor errors such as tense and plurality (Stubbs and Summerfield, 1990). Two different spatial configurations are considered: a two-source configuration at 0° and 5°, and a three-source configuration at −30°, 0°, and 30°. To eliminate potential location-based priming effects (Maljkovic and Nakayama, 1996) we randomly swap the locations for target and interference for different trials. In the unprocessed condition, binaural signals are produced by convolving original signals with the corresponding HRTFs and the convolved signals are presented to a listener dichotically (see Bodden, 1993). In the processed condition, our algorithm is used to reconstruct the target signal at the better ear and results are presented diotically.

Twelve native English speakers with normal hearing, between 24–30 years old, participated in the experiments. The tests were conducted in a sound-insulating booth (IAC model 40a-9) and signals were presented over Sennheiser HD 256 headphones. At the beginning of a test, subjects were familiarized with the voice of a target male speaker and they were free to adjust the sound volume to a comfortable level. The task of a subject during each test run was to report what was comprehended and a human operator marked the result. Each listener participated in a total of 8 conditions. Each condition contained 25 new, randomly chosen sentences, with the first 5 sentences used for practice only and their data discarded.

Figure 17 gives the keyword intelligibility score (median values and interquartile ranges) for the two-source configuration. Three SNR level are tested: 0, −5, and −10 dB, where the SNR is computed at the better ear for each sentence. The interfering source used for this configuration is babble noise. The general finding is that our algorithm improves the intelligibility score for the tested conditions. The improvement becomes larger as the SNR decreases (61% at −10 dB), even though the algorithm introduces more target distortions at lower SNR levels. Our informal observations suggest, as expected, that the intelligibility score improves for unprocessed mixtures when two sources are more widely separated than 5°. Figure 18 shows the results for the three-source configuration, where our model yields a 40% improvement. Here, the SNR is fixed at −10 dB at the better ear. The two interfering sources are one female speaker and a different male speaker. Note that, in this case, azimuth sepa-

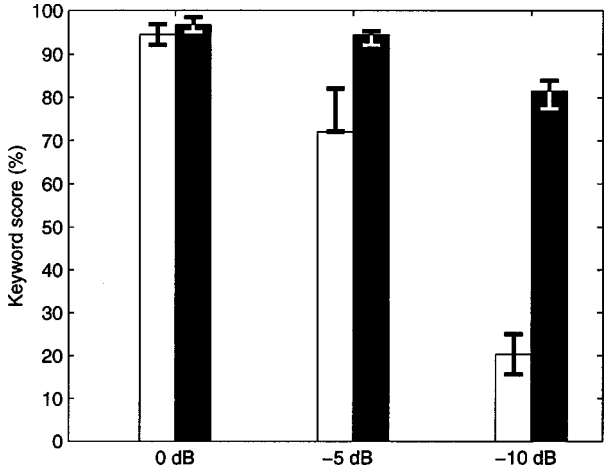

FIG. 17. Keyword intelligibility score (median values and interquartile ranges) before (white bars) and after processing (black bars) for a two-source condition (0° and 5°) at three SNR values: 0, −5, and −10 dB.

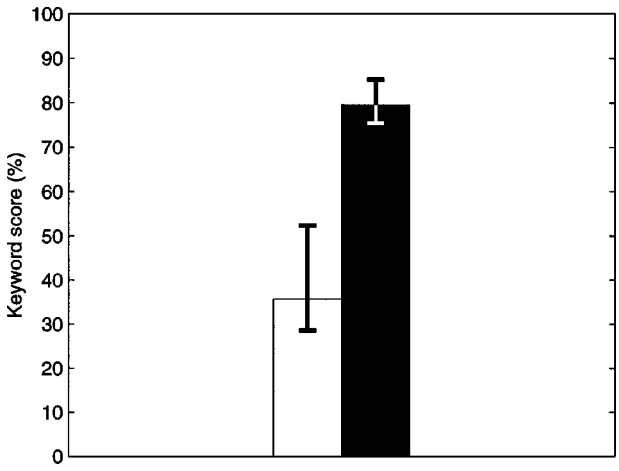

FIG. 18. Keyword intelligibility score (median values and interquartile ranges) before (white bars) and after processing (black bars) for a three-source condition (0°, 30°, and −30°) at −10-dB SNR.

ration is high between the three sources. Though we have not formally tested in the three-source configuration, we would expect that a trend similar to the one in Fig. 17 occurs with respect to SNR levels; that is, the model improvement decreases as SNR increases.

We recognize that comprehensive human subject evaluations of a model would require a separate study (e.g., see Stubbs and Summerfield, 1990), and indeed this is a topic we intend to pursue in the future. Nonetheless, as far as we know, our system is the first binaural model that has been shown to produce a large speech intelligibility improvement for normal listeners (see Kollmeier and Koch, 1994; Shamsoddini and Denbigh, 2001). The configurations and SNR conditions under which improvement occurs will be systematically characterized in future investigation.

VII. 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 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 sizable 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 through clustering have been proposed previously (Tessier and Berthommier, 1997; Lehn, 1997; Glotin *et al.*, 1999, Jourjine *et al.*, 2000). Notably, the experimental results by Jourjine *et al.* (2000) suggest that speech signals in a multiple-speaker condition obey to a large extent disjoint orthogonality in time and frequency. That is, at most one source has a nonzero energy at a specific time and frequency. Such models, however, assume input directly from microphone recordings and head-related filtering is 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. 13, 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 require that the number of sensors increases as the number of acoustic sources increases. A main reason for this difference is that considerations of human audition play a large role in our model design.

Conventional two-microphone adaptive beamformers can develop one deep null which cancels almost perfectly one interference under optimal conditions (Greenberg and Zurek, 2001). The performance, however, degrades when the number of interfering sources increases and is largely affected by the relative SNR of the individual interferences in the reference channel. Weiss (1987) measured the attenuation of individual interferences in acoustical mixtures across different conditions. The experimental results in the anechoic case show attenuation up to 14.5 dB in the two-source case, when both target and interference are active during filter adaptation. For the three-source case, the performance degrades across all interferences by 4 dB, and improvement can be as low as 0 dB. In comparison, our model works for a wide range of spatial configurations with two or more sources; for example, Fig. 13 shows that with three sources our model still obtains an average SNR gain of 11.3 dB. Conditions with high SNRs degrade the performance of adaptive beamforming. Our model, on the other hand, works especially

well for high-SNR scenarios. Subband versions of adaptive beamforming also exist (see, for example, Nordholm *et al.*, 2003). In this case, the signal is analyzed independently in frequency bands, and different directivity patterns are adaptively chosen in each band. This allows cancellation of multiple interferences with nonoverlapping spectra (Cezanne and Pong, 1995). Conventional adaptive beamformers with slow adaptation rate are unable to track fast spectral changes in a multispeaker scenario, resulting in suboptimal performance. Using a frame-by-frame multisource localization scheme, Liu *et al.* (2001) have proposed an equalization and cancellation system that has virtually zero adaptation time. Their two-microphone system exploits the location information in each frame and steers a different null in each frequency band, resulting in 6–7 dB gain in multispeaker scenarios. Our model uses a similar strategy, by employing the localization cue independently in each T–F unit in order to cancel simultaneous interferences. Hence, binaural processing models including ours may have advantages over adaptive beamformers in a range of acoustical situations.

In terms of limitations, our model currently does not address room reverberation or moving sound sources. Observe that supervised training is required for different spatial configurations. This limits the flexibility of our system to cope with, say, diffuse background noise. In addition, the localization of many sources in reverberant conditions with just two sensors is a challenging topic. The situation becomes more complex when source motion is considered. 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.*, 2003). Spatial and pitch information have both been utilized to simulate double-vowel recognition, showing added benefits for voiced stimuli (Lehn, 1997; Tessier and Berthommier, 1997). 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 multisource situation; 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 from the reference ones with respect to the relative strength between target and acoustic interference, and configuration-specific clustering in the joint ITD–IID feature space. Consequently, supervised learning of binaural patterns is 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 using both SNR and ASR measures. Evaluation results show that the system estimates ideal binary masks very well and performance degradation is gradual with increasing number and intensity of interferences. In addition, when tested with normal listeners, the model produces large speech intelligibility improvements for two-source and three-source conditions.

ACKNOWLEDGMENTS

The authors wish to thank the three anonymous reviewers for their constructive suggestions/criticisms. This research was supported in part by an AFOSR grant (F49620-01-1-0027) and an NSF grant (IIS-0081058). G.J.B. was supported by EPSRC grant GR/R47400/01. A preliminary version of this work is included in the Proceedings of 2002 ICASSP.

Arbib, M. A., editor (**2003**). *The Handbook of Brain Theory and Neural Networks*, 2nd ed. (MIT Press, Cambridge, MA).

Bench, J., and Bamford, J. (**1979**). *Speech Hearing Tests and the Spoken Language of Hearing-Impaired Children* (Academic, London).

Bernstein, L. R., and Trahiotis, C. (**2001**). ''Transposed stimuli reveal similar underlying sensitivity to interaural timing information at high and low frequencies,'' J. Acoust. Soc. Am. **109**, 2485–2485.

Blauert, J. (**1997**). *Spatial Hearing—The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA).

Bodden, M. (**1993**). ''Modeling human sound-source localization and the cocktail-party-effect,'' Acta Acust. (Beijing) **1**, 43–55.

Bodden, M. (**1996**). ''Auditory demonstrations of a cocktail party processor,'' Acustica **82**, 356–357.

Breebaart, J., van der Par, S., and Kohlrausch, A. (**2001**). ''Binaural processing model based on contralateral inhibition. I. Model structure,'' J. Acoust. Soc. Am. **110**, 1074–1088.

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

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

Brown, G. J., and Cooke, M. P. (**1994**). ''Computational auditory scene analysis,'' Comput. Speech Lang. **8**, 297–336.

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

Cezanne, J., and Pong, A. N. (**1995**). ''An adaptive subband differential microphone,'' J. Acoust. Soc. Am. **96**, 3262.

Colburn, H. S. (**1977**). ''Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise,'' J. Acoust. Soc. Am. **61**, 525–533.

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

Cooke, M. P., Green, P., Josifovski, L., and Vizinho, A. (**2001**). ''Robust automatic speech recognition with missing and unreliable acoustic data,'' Speech Commun. **34**, 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. **98**, 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. **102**, 2316–2324.

Darwin, C. J., and Hukin, R. W. (**1999**). ''Auditory objects of attention. The role of interaural time differences,'' J. Exp. Psychol. Hum. Percept. Perform. **25**, 617–629.

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

Durlach, N. I. (**1972**). ''Binaural signal detection: Equalization and cancellation theory,'' in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias (Academic, New York), Vol. II.

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

Gaik, W. (**1993**). ''Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling,'' J. Acoust. Soc. Am. **94**, 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,'' Hear. Res. **44**, 99–122.
Glotin, H., Berthommier, F., and Tessier, E. (**1999**). ''A CASA-labeling model using the localisation cue for, robust cocktail-party speech recognition,'' Proc. Eurospeech, Vol. 5, pp. 2351–2354.
Greenberg, J. E., and Zurek, P. M. (**2001**). ''Microphone—array hearing aids,'' in *Microphone Arrays: Signal Processing Techniques and Application*, edited by M. Brandstein and D. Ward (Springer, Berlin), pp. 229–253.
Hyväarinen, A., Karhunen, J., and Oja, E. (**2001**). *Independent Component Analysis* (Wiley, New York).
Jeffress, L. A. (**1948**). ''A place theory of sound localization,'' J. Comp. Physiol. Psychol. **41**, 35–39.
Jourjine, A., Rickard, S., and Yilmaz, O. (**2000**). ''Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures,'' in Proc. ICASSP, Vol. 5, pp. 2985–2988.
Kollmeier, B., and Koch, R. (**1994**). ''Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,'' J. Acoust. Soc. Am. **95**, 1593–1602.
Krim, H., and Viberg, M. (**1996**). ''Two decades of array signal processing research: The parametric approach,'' IEEE Signal Process. Mag. **13**, 67–94.
Lehn, K. H. (**1997**). ''Modeling binaural auditory scene analysis by a temporal fuzzy cluster analysis approach,'' Proc. IEEE WASPAA.
Leonard, R. G. (**1984**). ''A database for speaker-independent digit recognition,'' Proc. ICASSP, pp. 111–114.
Lindemann, W. (**1986**). ''Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation for lateralization for stationary signals,'' J. Acoust. Soc. Am. **80**, 1608–1622.
Liu, C., Wheeler, B. C., O'Brien, Jr., W. D., 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. **110**, 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. **39**, 604–622.
Maljkovic, V., and Nakayama, K. (**1996**). ''Priming of pop-out. II. Role of position,'' Percept. Psychophys. **58**, 977–991.
Moore, B. C. J. (**1997**). *An Introduction to the Psychology of Hearing*, 4th ed. (Academic, San Diego).
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. **45**, 224–240.
Nordholm, S. E., Claesson, I., and Grbic, N. (**2003**). ''Performance limits in subband beamforming,'' IEEE Trans. Speech Audio Process. **11**, 193–203.
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).
Popper, A. N., and Fay, R. R., editors (**1992**). *The Mammalian Auditory Pathway: Neurophysiology* (Springer, New York).
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. **89**, 807–817.
Shackleton, T. M., Meddis, R., and Hewitt, M. J. (**1992**). ''Across frequency integration in a model of lateralization,'' J. Acoust. Soc. Am. **91**, 2276–2279.
Shamsoddini, A., and Denbigh, P. N. (**2001**). ''A sound segregation algorithm for reverberant conditions,'' Speech Commun. **33**, 179–196.
Shinn-Cunningham, B. G., Schickler, J., Kopc-Carono, N., and Litovsky, R. (**2001**). ''Spatial unmasking of nearby speech sources in a simulated anechoic environment,'' J. Acoust. Soc. Am. **110**, 1118–1129.
Silverman, B. W. (**1986**). *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, New York).
Slatky, H. (**1993**). ''Algorithmen zur richtungsselektiven Verarbeitung von Schallsignalen eines binauralen Cocktail-Party-Prozessors,'' Ph.D. thesis, Ruhr-Universität Bochum.
Stern, R. M., and Colburn, H. S. (**1978**). ''Theory of binaural interaction based on auditory nerve data. IV. A model for subjective lateral position,'' J. Acoust. Soc. Am. **64**, 127–140.
Stern, R. M., and Trahiotis, C. (**1995**). ''Models of binaural interaction,'' in *Hearing*, edited by B. C. J. Moore (Academic, New York).
Stubbs, R. J., and Summerfield, Q. (**1990**). ''Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners,'' J. Acoust. Soc. Am. **87**, 359–372.
Tessier, E., and Berthommier, F. (**1997**). ''A model of the cumulative effect of pitch and interaural delay differences for double vowel segregation,'' Proc. ICSP Seoul.
Wang, D. L., and Brown, G. J. (**1999**). ''Separation of speech from interfering sounds based on oscillatory correlation,'' IEEE Trans. Neural Netw. **10**, 684–697.
Weintraub, M. (**1986**). ''A computational model for separating two simultaneous talkers,'' Proc. ICASSP, pp. 81–84.
Weiss, M. (**1987**). ''Use of an adaptive noise canceller as an input preprocessor for a hearing aid,'' J. Rehabil. Res. Dev. **24**, 93–102.
Whittkop, T., and Hohmann, V. (**2003**). ''Strategy-selective noise reduction for binaural digital hearing aids,'' Speech Commun. **39**, 111–138.
Wu, M., Wang, D. L., and Brown, G. J. (**2003**). ''A multipitch tracking algorithm for noisy speech,'' IEEE Trans. Speech Audio Process. **11**, 229–241.