LOCATION-BASED SOUND SEGREGATION

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

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

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 location-based approach for speech segregation. 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 region. We observe that within a narrow frequency band modifications to the relative energy of the target source with respect to the interfering energy trigger systematic deviations for binaural cues. 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 large improvement over previous models.

1. 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 [1] [2] [3]. The auditory system is able to segregate the speech signal from the acoustic mixture using various cues, including pitch, envelope, and location, in a process that is known as *auditory scene analysis* [1].

It is widely acknowledged that for human audition interaural time differences (ITD) represent the main binaural cue used at low frequencies (<2 kHz), whereas in the high frequency range both interaural intensity differences (IID) and interaural time differences between signal envelopes (IED) are used [2]. The resolution of the binaural cues has implications in both localization and recognition tasks. Experiments show that listeners can reliably detect 10-15 μs ITD from the median plane, which corresponds to 1-5 degrees in azimuth separation. On the other hand, the smallest detectable change in IID is about 0.5 dB to 1 dB at all frequencies. Resolution deteriorates as the reference azimuth gets larger and it has been reported to reach up to 10 degrees when the reference source is located far to the side of the head.

Increased speech intelligibility in binaural listening compared to the monaural case has prompted research in designing cocktail-party processors based on psychoacoustic principles [4] [5] [6]. In particular, building on a previous cross-correlation model for sound localization, Bodden 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 can produce substantial enhancement in speech intelligibility.

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 and the external ear. We introduce an *apriori* ideal binary mask that is motivated by the human auditory masking phenomenon, whereby the stronger signal masks the weaker one in the same critical band. If the original unmixed signals are available, one can construct the ideal mask in the following way: retain the time-frequency regions for which target energy exceeds interference energy and discard the other regions. 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 [7]. Hence, our model aims to estimate the ideal binary mask. Statistics for the relationship between the relative energy and the deviation of the binaural cues are at the core of our system. We show that for anechoic mixtures of multiple sound sources there exist strong correlation between the energy ratio and ITD and IID cues, resulting in a characteristic clustering. We employ a nonparametric classification method to determine decision regions for the ITD/IID features that correspond to an optimal estimate for the ideal mask.

Related models for estimating target masks have been proposed previously [8] [9]. Such models, however, assume input directly from microphone recordings. As a result, 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 in high frequencies than ITD. Second, frequency-dependent combinations of ITD and IID will arise naturally for a fixed spatial configuration. Consequently, channel-dependent training for each frequency band becomes necessary. Our tests with only ITD (as in [8]) or simple channel-independent classification (as in [9]) yield considerably inferior performance.

The rest of the paper is organized as follows. Section 2 describes the architecture of the model. Section 3 introduces a method for estimating the ideal binary mask. Section 4 presents systematic evaluation of the system for two and three sources and a comparison with the Bodden model.

0-7803-7402-9/02/$17.00 ©2002 IEEE

2. MODEL ARCHITECTURE

Our model consists of the following four stages: 1) a physiological model of auditory periphery; 2) binaural cue extraction; 3) azimuth localization for both target and interferences; and 4) estimation of the ideal binary mask.

The input to our model is a mixture of two or more signals presented at different, but fixed, locations: target speech and acoustic interfererences, which are sampled at 44.1 kHz. Binaural signals are obtained by convolving the input with measured head related impulse responses (HRIR) from a KEMAR dummy head [10].

To simulate the auditory periphery we use a bank of 128 gammatone filters in the range of 80 Hz to 5 kHz as described in [11]. In addition, the gains of the gammatone filters are adjusted in order to simulate the middle ear transfer function. In the final stage of the peripheral model, the output of each gammatone filter is half-wave rectified in order to simulate the firing probabilities of the auditory nerve. To process the envelopes of the signals at higher frequencies a low-pass filter with cutoff frequency of 800 Hz is used. Saturation effects are modeled by taking the square root of the signal.

Current models of azimuth localization almost invariably start with the cross-correlation mechanism. We utilize the following azimuth localization method:

- Compute interaural cross-correlation coefficients at time delays equally distributed in the plausible range from –1 ms to 1 ms for all frequency channels.
- In a training phase, we derive frequency-dependent nonlinear transformations and map the time-delay axis onto an azimuth axis.
- Calculate an energy-weighted summary across time and frequency and identify the peaks, which generally are very close to the true source locations.

3. BINARY MASK ESTIMATION

Our objective is to develop an efficient algorithm for estimating the ideal binary mask. Our estimation is based on the following observation regarding the acoustic interaction of multiple sources: in a narrow band, the ITD and IID corresponding to the target source centers around azimuth-dependent characteristic values. As the interference from additional sound sources increases, ITD and IID systematically deviate from these values.

3.1 Pure Tones

In order to investigate the relationship between relative signal strengths and ITD for two pure tones with the same frequency ω we derive the cross-correlation function for the mixture:

$$c(\tau) = \frac{A_1^2}{2}\cos(\omega(\tau - d_1)) + \frac{A_2^2}{2}\cos(\omega(\tau - d_2)) + \\ + \frac{A_1 A_2}{2}\cos(\omega(\tau - \frac{d_1 + d_2}{2}))\cdot coeff \quad (1)$$

in which A_i is the amplitude, d_i represents the time delay for the ith source, and $coeff = \cos(\Delta\varphi + \omega(d_2 - d_1)/2)$ where $\Delta\varphi$ is a function of phase differences between the initial signals and those due to the arrival times of the signals at the left ear. For simplification, IID is considered negligible – true for low-frequency channels. By observing deviation of the peak location $\tau_{\max}$ from the middle location $(d_1 + d_2)/2$, we obtain which source is stronger.

On the other hand, we estimate IID as the ratio of energy at the two ears. Therefore, for two pure tones we have:

$$\mathrm{IID} = 10\log_{10}\frac{A_1^2\cdot|H_1^r(\omega)|^2 + A_2^2\cdot|H_2^r(\omega)|^2}{A_1^2\cdot|H_1^l(\omega)|^2 + A_2^2\cdot|H_2^l(\omega)|^2} \quad (2)$$

where $H_i^r(\omega)$ and $H_i^l(\omega)$ represent the right/left HRTF for the ith source.

A systematic change in the relative amplitude results in a corresponding shift for both ITD and IID. Moreover, thresholds can be derived in order to decide which of the two signals is stronger in a specific region.

3.2 Real signals

Our model treats ITD and IID as two separate dimensions in the feature space. This integration of ITD and IID exploits the best discrimination power of the two binaural cues for a specific configuration.

We estimate independently for all frequency channels the local ITD and IID and the energy ratio E based on 20-ms time frames with 10 ms overlap between adjacent time frames. IID and energy ratios for the ith channel are computed as follows:

$$L_i = 20\log_{10}\sum_t l_i^2(t) \Big/ \sum_t r_i^2(t) \quad (3)$$

$$E_i = \sum_t s_i^2(t) \Big/ (\sum_t s_i^2(t) + \sum_t n_i^2(t)) \quad (4)$$

where l_i and r_i refer to the left and right auditory periphery output of the ith channel, respectively; s_i refers to the output for the target signal, whereas n_i corresponds to the acoustic interference. In computing IID, we use 20 instead of 10 in order to account for the square root operation in peripheral processing.

In order to eliminate the multiple peaks in the cross-correlation function for mid- and high-frequency channels, we consider the following strategy. We compute the ideal ITD for the target source, loc_i for the ith channel. We study deviations from loc_i due to the interferences from other sources. Consequently, a local ITD is estimated as the delay $\tau_i^{\max}$ that corresponds to maximum activity in the cross-correlation pattern in the range $[loc_i - \pi, loc_i + \pi]$.

ITD and IID undergo relatively smooth changes with the energy ratio in a given frequency channel. To capture this relationship, statistics are collected for a given spatial configuration. We employ the corpus collected by Cooke [10], which is commonly used in sound separation studies. The corpus has 100 mixtures obtained from 10 speech utterances mixed with 10 noise intrusions, encompassing a variety of common acoustic interferences such as telephone ringing, rock music, and other speech. Half of the corpus is used for training, and testing is done on the rest of the corpus. Fig. 1 displays statistics for a channel with center frequency about 1.5 kHz obtained when

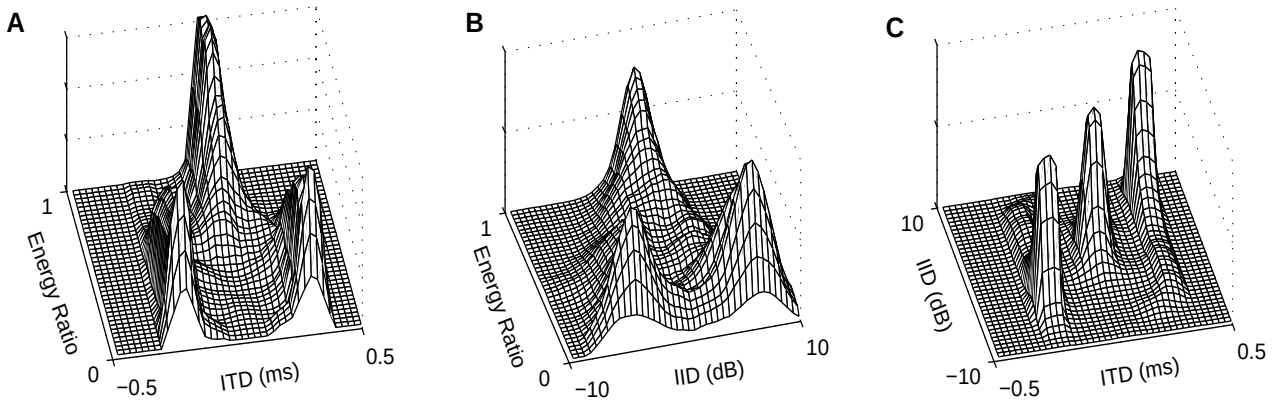

Figure 1: A: Relationship between ITD and the energy ratio. B: Relationship between IID and the energy ratio. C: Clustering in the ITD-IID feature space. Histograms are obtained for speech at 0° and interfering noise at 30° and -30°, for the channel number 85 with center frequency ~1.7 kHz.

the target is presented at 0° (median plane) and two other sources are active, at -30° and 30°. When the information is displayed in the ITD-IID plane, we observe location-based clustering of the binaural cues (Fig. 1C).

Since we are interested in estimating a binary mask, we focus on detecting decision regions in the 2-dimensional ITD-IID feature space. Consequently, standard supervised learning techniques can be applied. For the *i*th channel, we test the following two hypotheses. The first one is H_0: target is dominant or $E_i > 0.5$, and the second one is H_1: interference is dominant or $E_i < 0.5$. Based on estimates of the bivariate densities $p(x|H_0)$ and $p(x|H_1)$ the classification is done in accordance with the maximum a posteriori decision rule:

$$\delta(x) = \begin{cases} 1, & \text{if } p(H_0)p(x|H_0) > p(H_1)p(x|H_1) \\ 0, & \text{else} \end{cases} \quad (5)$$

There exist a plethora of techniques for probability density estimation ranging from parametric techniques (e.g. mixture of gaussians) to nonparametric techniques (e.g. kernel density estimators). In order to completely characterize the distribution of the data we decided to use a kernel density estimation method. 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 it does not generalize well, for too large values the structure of the data distribution will disappear. One approach for finding the optimal values is the least squares cross validation method [12], which is utilized in our estimation. Optimal values of the parameters are chosen as local minima in the range $[1/4n^{-1/6}\sigma_i,\ 3/2n^{-1/6}\sigma_i]$ where σ_i represents the variance for the *i*th smoothing parameter and *n* is the sample data number.

4. PERFORMANCE AND COMPARISON

In order to evaluate the performance of the system for speech segregation, the segregated signal is reconstructed from an estimated binary mask following a resynthesis method described by Brown and Cooke [13]. In order to quantitatively assess system performance, we measure in decibels the SNR using the original speech before mixing as signal:

$$SNR = 10\log_{10} \sum_t s_o^2(t) \Big/ \sum_t \left(s_o(t) - s_e(t)\right)^2 \quad (6)$$

where $s_o(t)$ represents the resynthesized original speech signal and $s_e(t)$ the reconstructed speech from an estimated binary mask signal. One can similarly measure the SNR of the mixture by replacing the denominator with $s_N(t)$, the resynthesized original interference.

We compare the SNR gain against results obtained using the ideal binary mask. For two-source segregation, the system is systematically evaluated at the "better ear" for various combinations of azimuth angles. When the target is in the median plane, excellent results are obtained for azimuth separation as small as 5°, as shown in Fig. 2A. Performance degrades when the target source is moved to the side of the head, as shown in Figs. 2B and 2C. This pattern of performance is in agreement with psychoacoustic data [2]. When comparing the SNR with the SNR in the initial mixture, there is an average SNR gain of 14 dB for target sources in the median plane. The gain reduces to 11 dB when the target source is at 70°.

Our approach can be extended to cases with more than two sources. Localization methods based on cross-correlation, including ours, are not limited to two locations. With identified locations, our model performs target segregation in a similar manner. Fig. 2D illustrates the performance of our model for the case of three sources with target located in the median plane and two interfering sources at –30° and 30°. The average SNR gain obtained is approximately 10 dB. This property of our model differs from many blind source separation and array processing methods where the number of sensors must be no smaller than the number of sources.

In order to draw a quantitative comparison, we have implemented the Bodden model [6], which produces good sound separation using source locations. First, we note that Bodden's cocktail-party processor is a great deal more complicated than ours. His system uses a 24-channel filterbank intended to simulate critical bands and an extended cross-correlation mechanism based on contralateral inhibition in order to compute

ITD in the low-frequency range and IID for the high-frequency range. For a fair comparison, our implementation of the Bodden system uses the same 128 channel gammatone filterbank; we also implemented the 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, and our choice of (-10°, 30°) falls into the range where his model performs optimally. As shown in Fig. 3, our model shows a considerable improvement ~3 dBover the Bodden system (~3 dB).

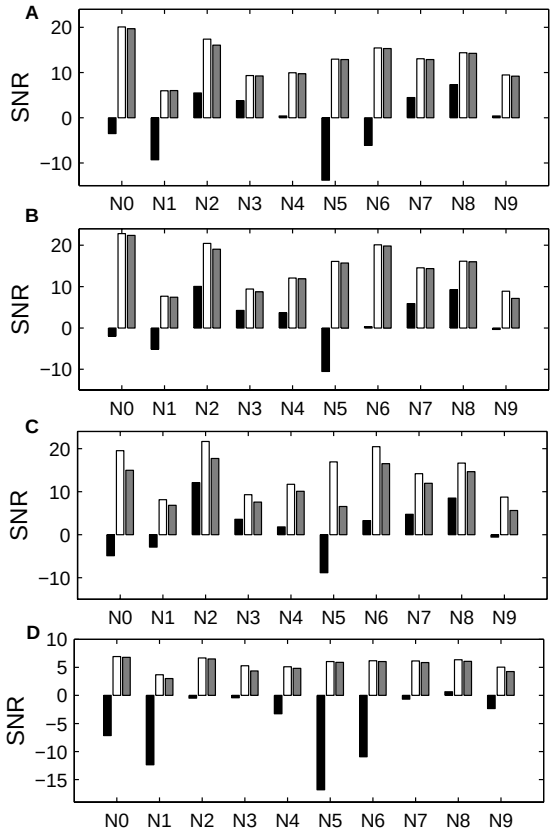

Figure 2: Systematic evaluation. Black bar represents the mixture SNR, white bar is the SNR using the ideal binary mask and gray bar corresponds to results from our system. The corpus contains 10 male utterances mixed with ten interferences (N0: pure tone; N1: white noise; N2: noise burst; N3: 'cocktail party'; N4: rock music; N5: siren; N6: trill telephone; N7: female speech; N8: male speech; N9: female speech). A: Target at 0°, interference at 5°. B: Target at 45°, interference at 50°. C: Target at 70°, interference at 75°. D: three-source configuration: target at 0°, interference at -30° and 30°.

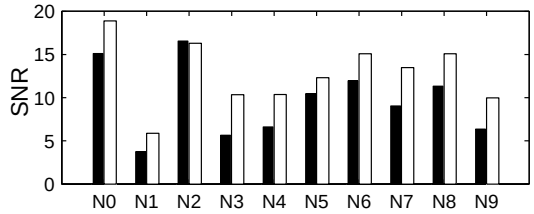

Figure 3: SNR comparison between the Bodden model (black bar) and our model (white bar). Target is at 30° and interference -10°.

5. CONCLUSION

We have presented a location-based sound segregation system, motivated by psychoacoustic and physiological studies of the auditory system. The input to the system is obtained by convolving the original signals with direction-dependent HRIRs. The system can be applied to spatial configurations with two or more sources. Our approach is based on an analysis of the relationship between ITD/IID and target/interference energy ratio within narrow bands. Our model yields segregation results that constitute a significant improvement over previous models.

Our study proves that the binaural cues are very effective in filtering out acoustical interference in anechoic room conditions. We are currently extending our model to deal with room reverberations by incorporating the precedence effect and forward/backward masking.

Acknowledgements

This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-0027).

6. REFERENCES

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

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

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

[4] R. F. Lyon, "A computational model of binaural localization and separation," *Proc. IEEE ICASSP*, 1983.

[5] J. Lazarro and C. Mead, "A silicon model for auditory localization," *Neural Computation*, vol. 1, pp. 47-57, 1989.

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

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

[8] H. Glotin, F. Berthommier and E. Tessier, "A CASA-Labelling model using the localisation cue for, robust cocktail-party speech recognition," *Proc. EUROSPEECH*, pp.2351-2354, 1999.

[9] A. Jourjine, S. Rickard and O. Yilmaz, "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures," *Proc. IEEE ICASSP*, vol. 5, pp. 2985-2988, 2000.

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

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

[12] B. W. Silverman, *Density estimation for statistics and data analysis*, New York: Chapman and Hall, 1986.

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