A BAYES-RULE BASED HIERARCHICAL SYSTEM FOR BINAURAL SOUND SOURCE LOCALIZATION

Danfeng Li and Stephen E. Levinson

University of Illinois at Urbana-Champaign
Dept. of Electrical and Computer Engineering
405 N. Mathews Ave., Urbana, IL 61801, USA

ABSTRACT

A Bayes-rule based hierarchical binaural sound source localization system is proposed. By combining three localization cues: interaural time differences (ITDs), interaural intensity differences (IIDs), and spectral cues, and a hierarchical decision making structure, this system enables to locate a sound source in a 3D space by using only the binaural inputs. Preliminary simulations have shown the effectiveness of this system. It can be used in studying binaural localization mechanism and applications such as in hearing aids and robotics.

1. INTRODUCTION

The study of binaural localization mechanism can be traced back to the late 19th century, when British physicist Lord Rayleigh proposed the "Duplex Theory" for localization in the lateral dimension [1]. The "Duplex Theory" states that interaural time differences (ITDs) provide cues to the lateral positions of low-frequency sounds, and interaural intensity differences (IIDs) provide cues of high-frequency sounds. Rayleigh's theory has been proven in numerous psychophysical and physiological studies, but failed to explain the front/back discrimination and vertical localization by using these binaural cues. Research in the past few decades has revealed the importance of monaural cues, which are so far mainly spectral cues, for localization.

Most of the existing binaural models use only ITDs for sound source localization. They are all based on the coincidence model proposed by Jeffress in 1948 [2], which is basically a cross-correlation calculator to find the time difference of arrival between two ears. A few efforts have been made to use IIDs and spectral cues in localization models. Fuzessery used diagrams and the differences of sound intensity levels at three arbitrary frequencies, by assuming flat spectrum of a sound source [3]. Zakarauskas and Cynader further improved Fuzessery's work by relaxing the flat spectrum assumption [4]. They proposed a computational model using spectral cues for localization by applying the first and second order finite differences of sub-band intensities of the spectrum, assuming the source spectrum has an either locally constant intensity or constant slope. Neti *et al.* proposed a three-layer neural network for localization by feeding ITDs, IIDs and the spectrums from both channels into the network [5].

In this paper, a Bayes-rule based hierarchical binaural sound source localization system is proposed, which employs all ITDs, IIDs and spectral cues. The decision is made in three steps. First, a set of possible locations are selected by ITDs. Then these locations are further narrowed down by IIDs. The final decision is made by spectral cues. It will be explained later, this is a plausible approach from both psychophysical and engineering point of views.

2. LOCALIZATION CUES EXTRACTION

Denote the source signal as $s(n)$, and the received signals at the left and the right ear as $x_l(n)$ and $x_r(n)$, respectively. The following relations hold.

$$\begin{aligned} x_l(n) &= h_l(n) * s(n) + n_l(n) \\ x_r(n) &= h_r(n) * s(n) + n_r(n) \end{aligned} \tag{1}$$

Where $h_l(n)$ and $h_r(n)$ are the transfer functions for the direct paths to the two ears, and $n_l(n)$ and $n_r(n)$ are the received noises. In general, the noises consist two parts, one is due to the reverberation, which is related to the source $s(n)$, another is due to other sources in the environment, which can be considered independent with the dominant source $s(n)$. To emphasis the key concepts of the localization cues extraction scheme, the effects of noise and reverberation are ignored in this paper.

2.1. Interaural Time Differences (ITDs)

If further assume that each transfer functions $h_l(n)$ and $h_r(n)$ can be approximated by a pure magnitude decay and a pure time delay, the cross-correlation between the signals from

two ears become

$$R_{x_l x_r} = \alpha R_{ss}(\tau - D). \tag{2}$$

Where α is a scalar and D is the interaural time difference (ITD). This assumption holds for the low frequency components, where the main interaural difference is the time of arrival. The cross power spectrum is defined as the Fourier transform of the cross correlation of the two signals

$$G_{x_l x_r}(\omega) \equiv \sum_{n=0}^{N} R_{x_l x_r}(\tau) e^{j\omega\tau}. \tag{3}$$

It can be calculated as

$$G_{x_l x_r}(\omega) = X_l(\omega) X_r^*(\omega). \tag{4}$$

where the superscription $*$ denotes the complex conjugate. Take the Fourier transform of (2), the cross power spectrum of the received signals is

$$G_{x_l x_r}(\omega) = \alpha G_{ss}(\omega) e^{-j\omega D}. \tag{5}$$

It indicates that the ITD D is only related to the phase of the cross power spectrum. The normalized cross-correlation [6] is given by

$$\begin{aligned} \hat{R}_{x_l x_r}(\tau) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{G_{x_l x_r}(\omega)}{|G_{x_l x_r}(\omega)|} e^{j\omega\tau} d\omega \\ &= \delta(\tau - D). \end{aligned} \tag{6}$$

Therefore the ITD D can be estimated as

$$D = \arg\max_{\tau} \hat{R}_{x_l x_r}(\tau). \tag{7}$$

2.2. Interaural Intensity Differences (IIDs) and Spectral Cues

The extractions of IIDs and spectral cues share many common processes. First, both of them are calculated in intensity domain. This is due to two reasons. One is from a neuroscience point of view, the response of neurons in the auditory system is roughly proportional to intensity levels [7]. The other is from an engineering point of view, in intensity domain, the relation between signal and channel transfer function becomes simple addition. Secondly, both of them are extracted for each sub-band instead of the whole spectrum as a whole, and for the high-frequency sub-bands only. This is because the detailed shape of the spectrum can provide useful information for localization, and the intensity difference is negligible at the low-frequency sub-bands. The diagram of IIDs and spectral cues extraction is shown in Fig. 1.

If ignoring the background and reverberation noises, the power spectrums of the received signals at two ears are

$$\begin{aligned} P_l(\omega) &= S(\omega)|H_l(\omega)|^2 \\ P_r(\omega) &= S(\omega)|H_r(\omega)|^2, \end{aligned} \tag{8}$$

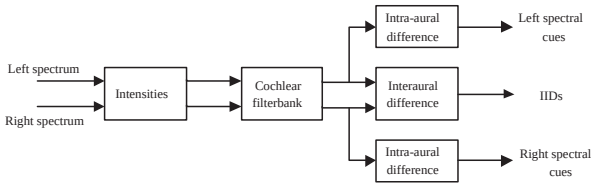

Fig. 1. Diagram for IIDs and spectral cues extraction.

where $S(\omega)$ is the power spectrum of the source and $H_l(\omega)$ and $H_r(\omega)$ are the transfer functions to the two ears. The intensities are

$$\begin{aligned} I_l(\omega) &= 10\log P_l(\omega) = 10\log S(\omega) + 20\log|H_l(\omega)| \\ I_r(\omega) &= 10\log P_r(\omega) = 10\log S(\omega) + 20\log|H_r(\omega)| \end{aligned} \tag{9}$$

After the intensities have been calculated, the signals then are passed through a cochlear filterbank. For each qualified sub-band i, the following relations hold,

$$\begin{aligned} I_i &= \int_{\omega\in\Omega_i} W_i(\omega)[10\log S(\omega) + 20\log H(\omega)]d\omega \\ &= 10\int_{\omega\in\Omega_i} W_i(\omega)\log S(\omega)d\omega \\ &\quad + 20\int_{\omega\in\Omega_i} W_i(\omega)\log H(\omega)d\omega \\ &\equiv \chi_i + F_i, \end{aligned} \tag{10}$$

where Ω_i is the frequency range for sub-band i, and $W_i(\omega)$ is the weight from the cochlear filter. The subscriptions for the left and the right channels are dropped to simplify the notation.

2.2.1. IIDs

The IID for sub-band i is defined as

$$\begin{aligned} E_i &\equiv I_{li} - I_{ri} \\ &= (\chi_i + F_{li}) - (\chi_i + F_{ri}) \\ &= F_{li} - F_{ri}. \end{aligned} \tag{11}$$

Because the signal intensity χ_i is the same for the left and the right channel, IIDs do not depend on the input signal in general. However, for a narrow-band signal, IID will be zero at the subband where the signal power is zero. Therefore, IIDs will become signal depended. This problem can be solved by accumulating IIDs over time, if assuming the signal is not always narrow-band.

2.2.2. Spectral Cues

The use of spectral cues in this paper mainly follows the work done by Zakarauskas *et al.* [4], where the first and the second finite differences of a spectrum are applied.

As shown in (9), the intensity of the received signal at each ear can be separated into two independent parts. One is related only to the signal and the other is related only to the transfer function. Assuming that the signal log power spectrum is relative flat, $\chi_i \approx \chi_{i+1}$, and the first finite difference between the observed intensity levels I_i and I_{i+1} is

$$\begin{aligned} D_i &\equiv I_{i+1} - I_i \\ &= (\chi_{i+1} - \chi_i) + (F_{i+1} - F_i) \\ &\approx F_{i+1} - F_i, \qquad i = 1, N-1, \end{aligned} \tag{12}$$

which is only related to the transfer function. If the condition $\chi_i \approx \chi_{i+1}$ does not hold, but the slope of the spectrum changes slowly, the second finite difference can be used

$$\begin{aligned} C_i &\equiv D_{i+1} - D_i \\ &\approx F_{i+2} - 2F_{i+1} - F_i, \qquad i = 1, N-1. \end{aligned} \tag{13}$$

The assumptions for (12) and (13) are valid for some most often observed signals, such as speech. It is mainly due to the applying of the cochlear filterbank. The cochlear filterbank intends to distribute the energy equally among its sub-bands for the most observed signals. The spectral cues can be further calculated by using several frames instead of one. The accumulated log power spectrum intend to be more flat, which makes the assumptions more valid.

Note this approach is a little different with Zakarauskas's approach, where filterbank is applied before the calculation of intensities [4]. In Zakarauskas's approach, (10) only approximately holds by assuming the local spectrum is relatively flat. Here, the relation holds exactly without any assumption, therefore results in more accurate IIDs and spectral cues estimation. The comparison between the two methods will be provided in the simulations.

3. DECISION MAKING

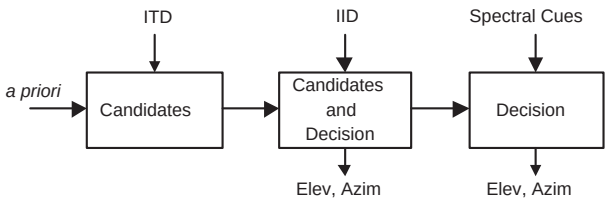

Fig. 2. Bayes-rule based hierarchical decision making.

A Bayes-rule based hierarchical decision making approach is employed. It is shown in Fig. 2. The location probability is first calculated by ITDs and the *a priori* information. This probability servers as the *a priori* for the second step, where the location probability is refined by IIDs. The refined probability is in turn served as the *a priori* for the last step, and the spectral cues are combined to make the final decision. The reasons for such an approach are both from psychophysical evidences and engineering consideration. It has be shown by human listener experiments that the ITDs are the most robust localization cues, and then the IIDs, and the spectral cues are very sensitive to the environment and input signals [8]. In term of signal processing, ITDs can be more accurately estimated comparing to IIDs, and again, spectral cues are still related to the input signals even after the processes described in the previous section.

Mathematically, the decision procedure can be expressed as

$$\begin{aligned} P(\xi|\mathbf{o}) &= P(\xi|\mathbf{o}_{itd}, \mathbf{o}_{iid}, \mathbf{o}_{spe}) \\ &= \frac{P(\mathbf{o}_{spe}|\xi, \mathbf{o}_{itd}, \mathbf{o}_{iid}) P(\xi|\mathbf{o}_{itd}, \mathbf{o}_{iid})}{P(\mathbf{o}_{spe}|\mathbf{o}_{itd}, \mathbf{o}_{iid})} \\ &= \frac{P(\mathbf{o}_{spe}|\xi)}{P(\mathbf{o}_{spe})} \frac{P(\mathbf{o}_{iid}|\xi, \mathbf{o}_{itd}) P(\xi|\mathbf{o}_{itd})}{P(\mathbf{o}_{iid}|\mathbf{o}_{itd})} \\ &= \frac{P(\mathbf{o}_{spe}|\xi)}{P(\mathbf{o}_{spe})} \frac{P(\mathbf{o}_{iid}|\xi)}{P(\mathbf{o}_{iid})} \frac{P(\mathbf{o}_{itd}|\xi) P(\xi)}{P(\mathbf{o}_{itd})} \end{aligned} \tag{14}$$

where $P(\xi|\mathbf{o})$ is the probability of the source locates at ξ given the observation $\mathbf{o}$; $P(\mathbf{o}_{itd}|\xi)$, $P(\mathbf{o}_{iid}|\xi)$ and $P(\mathbf{o}_{spe}|\xi)$ are the observation probabilities for ITDs, IIDs and spectral cues given the source location, and $P(\xi)$ is the *a priori* information. The simplification of the expression is base on the independence assumption between localization cues. The decision rule therefore is

$$\begin{aligned} \xi &= \arg\max_{\xi} P(\xi|\mathbf{o}) \\ &= \arg\max_{\xi} \{P(\mathbf{o}_{spe}|\xi) P(\mathbf{o}_{iid}|\xi) P(\mathbf{o}_{itd}|\xi) P(\xi)\}. \end{aligned} \tag{15}$$

It should be noticed from Fig. 2 that the finally decision of the source location could also be made at the second step. As we already knew that the ITDs and IIDs can provide discrimination for most of the locations except for the median plane, and they are more robust than spectral cues. Therefore, the spectral cues are only used for locating sources in the median plane, where ITDs and IIDs provide no information.

In order to apply this decision making procedure, the 3D space need to be divided into regions according to the elevation and azimuth. For each individual region, the true ITDs, IIDs and spectral cues are found either through HRTF measurement, if it is available, or throught training, for most of the cases. During the training, ITDs, IIDs and spectral cues are the averages across the time from the training tokens.

Donate the feature spaces of IIDs, spectral cues for the left and right channel as

$$\begin{aligned} \mathbf{E} &= [E_1, E_2, \cdots, E_N] \\ \mathbf{C_l} &= [C_{l1}, C_{l2}, \cdots, C_{lM}] \\ \mathbf{C_r} &= [C_{r1}, C_{r2}, \cdots, C_{rM}], \end{aligned} \tag{16}$$

where E_i is the IID, and C_{li}/C_{ri} is the 2nd order difference for each sub-band as discribed in the previous section.

The cosine distance is applied for the similarity measurement, take the IIDs as an example

$$\alpha_E = \frac{<\bar{\mathbf{E}}, \mathbf{E}>}{||\bar{\mathbf{E}}||||\mathbf{E}||} \tag{17}$$

where $<,>$ denotes inner product, $||\cdot||$ denotes the 2nd order norm and $\bar{\mathbf{E}}$ is true IID.

4. SIMULATIONS

Simulations are conducted using pre-measured HRTF and pre-recorded audio data. The HRTF is measured by Gardner and Martin in MIT Media Laboratory [9]. Three data sets are used in simulations, commands (COM), cnn (CNN) and authentic sound effects (ASE). The COM data set contains 16 commands, such as "back", "close", "open", etc. The CNN data set contains about 5 hours recording from channel cnn headline news, it contains human speech, music, mixed speech and music, and other sounds. The ASE is from the four-volume CD set produced by Keith Holzman, it contains human noise, machine noise and other nature sounds. These three data sets cover many most often heard sounds in our daily life.

In the simulations, the 3D space is divided into 3 elevations, $0°$, $30°$ and $60°$. Assuming that the ITD estimation is always correct (it is always true without noise and reverberation), the decision for the left and right will therefore always correct. Under this assumption, only half of the space need to be considered. For the elevation at $0°$, the azimuth is equally divided into 18 regions, $10°$ each, the elevation at $30°$ is divided into 15 regions, $12°$ each, and the elevation at $60°$ into 9 regions, $20°$ each.

Total five sets of simulation are conducted. The error percentages are shown in Table 1. In the first simulation, DIR, the "ture" center for each location is calculated directly from the HRTF. In the second simulation, TRA, the center for each location is trained by the command set (COM). It shows the efficiency of the extractions of IIDs and spectral cues. In the third simulation, TOG, the IIDs and spectral cues are used together instead of in the hierarchical manner, which shows the advantage of using the hierarchical decision making procedure. The fourth simulation, MFCC, shows the results of applying filterbank on power spectrum instead of intensity as as mentioned in Section 2.2.2. The last simulation, MON, shows the localization ability of using monaural cues (spectral cues) only.

5. CONCLUSION AND DISCUSSION

A Bayes-rule based hierarchical binaural sound source localization system is proposed. The system employs two binaural cues, ITDs and IIDs, and a monoaural cue, the spectral cues, to localize a souce in a 3D space. Preliminary simulation results show the effectiveness of the algorithm.

Further research will be conducted on the effect of noises and reverberations, and the building of the *a priori* by combining source location history and inputs from other modalities, such as vision.

Table 1. Localization Error percentages for different schemes

Data	DIR	TRA	TOG	MFCC	MON
COM	0.0	0.0	0.4	0.6	20.0
CNN	0.0	0.0	2.2	5.9	37.6
ASE	0.4	0.7	2.6	10.4	38.7

6. REFERENCES

[1] W. J. Strutt, "On our perception of sound direction," *Philos. Mag.*, vol. 13, pp. 214–232, 1907.

[2] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psych.*, vol. 61, pp. 468–486, 1948.

[3] Z. M. Fuzessery, "Speculations on the role of frequency in sound localization," *Brain Behav. Evol.*, vol. 28, 1986.

[4] P. Zakarauskas and M.S. Cynader, "A computational theory of spectral cue localization," *J. Acoust. Soc. Am.*, vol. 94, no. 3, pp. 1323–1331, Sep. 1993.

[5] C. Neti, E. D. Young, and M. H. Schneider, "Neural network models of sound localization based on directional filtering by the pinna," *J. Acoust. Soc. Am.*, vol. 92, pp. 3140–3156, 1992.

[6] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[7] M. B. Sachs and P. J. Abbas, "Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli," *J. Acoust. Soc. Am.*, vol. 56, pp. 1835–1847, 1974.

[8] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 2219–2236, May 2002.

[9] B. Gardner and K. Martin, "Hrtf measurements of a kemar dummy-head microphone," Tech. Rep. 280, MIT Media Lab, May 1994.