Binaural Sound Source Localization Based on Steered Beamformer with Spherical Scatterer

Zhao Shuo, Chen Xun, Hao Xiaohui, Wu Rongbin, Wu Xihong

National Laboratory on Machine Perception, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. 100871.

Correspondence should be addressed to Wu Xihong (wxh@cis.pku.edu.cn)

ABSTRACT

Inspired by human sound localization, this paper introduces a novel approach to achieving binaural sound source localization in the frontal azimuthal half-plane based on steered beamformer. Instead of using HRTFs, a rigid sphere, whose transfer functions can be calculated accurately, is introduced to simulate the head effect. Sub-band beamformer using both the time cue ITD and the intensity cue IID is designed to process sound scattered by the rigid sphere. In the multi-band processing, a specialized filterbank is designed and the joint judgement strategy is employed. The whole system consists of the algorithm part, the hardware part and the user interface. Evaluation results by simulation and measurement experiments are also presented.

1. INTRODUCTION

Sound source localization is defined as the determination of the coordinates of sound sources in relation to a point in space. Many automation systems, such as voice capturing in conference room, hearing aids and security monitor, require sound source localization. Over the last two decades, many approaches are proposed to solve this problem, such as time-delay estimation (TDE) [1, 2, 3], beamforming [4, 5, 6, 20], hemisphere sampling [7], and accumulated correlation [8, 9]. Most of these methods rely exclusively upon the “time cue”, which is known as the interaural time difference (ITD) or the interaural intensity difference (IID). But due to the Duplex theory[10], there is another important cue for sound source localization, the interaural intensity difference (IID) or the interaural level difference(ILD), which has received little attention in the signal processing community[11]. In recent years, more and more localization algorithms are applied in robotic localization and many robotic localizing devices have appeared. In these applications, there are mainly three types of systems, including sensor arrays in free field, sensor arrays on robots without considering intensity information[20] and sensor arrays on robots considering both time and intensity cues[16, 19]. The last type which tries to mimic human audition has become more and more popular recently.

The ability of humans to perceive the spatial location of sound in space is truly a remarkable skill. In nature, human brain uses interaural differences in various frequency bands to infer the location of a source[12, 13]. Both cues mentioned above for sound source localization are included in Head-Related Transfer Functions (HRTFs), \(HRTF_{LR}\), which are the function of sound source location, frequency, and also the human subject. The HRTFs, which characterize the scattering properties of a person’s anatomy (especially the head, pinnae and torso), play a very important role in binaural source localization for human. Due to the scattering of the anatomy, the IID cue becomes more notable compared to the free sound field situation and both cues vary with frequency, which brings more information for the sound source localization task.

Inspired by human binaural sound localization, we are interested in developing some artificial binaural sound source localization system with a human-like scatterer. It seems that dummy-head is a suitable choice for the localization task because it has complex structures to approach the human head. However, there are lots of limitations using the dummy head. First of all, HRTFs of dummy head are usually obtained by measurement, which is both time-consuming and experimentally difficult to get the results accurately[14]. Low-frequency measurements are particularly problematic, partly be-
cause large loudspeakers are required, and partly because even good anechoic chambers reflect long wavelength sound waves. Although the numerical computation can bring more accurate results, it requires a huge project including scanning, modeling, meshing, long-time computation and post-processing. Second, the dummy head does not fit in every application situation because of its "big body" especially when the localization part is required to be integrated in a bigger system. Another method which does not require HRTFs is the auditory epipolar geometry which can extract directional information of sound sources without using HRTFs[15]. But it judges only three directions - left, right or center - for IID, while it gives continuous estimation of IPD against a direction of a sound source[16].

Instead of the dummy-head, a rigid sphere with two microphones is employed here to generate the binaural cues because its transfer functions are fairly similar to HRTFs in low and medium frequencies, in which most of the speech energy is contained. The Scattering Theory[17] is used here which is concerned with the effect obstacles or inhomogeneities on incident waves. It provides a method to estimate the scattered field from the knowledge of the incident field and the scattering obstacle. Transfer functions of a rigid sphere can be computed very easily due to the analytical solution of its scattered sound field using the scatterer theory[18], so that the binaural cues - ITD and IID - can be accurately obtained.

In previous studies, [16, 19] also employed the rigid sphere with two microphones to simulate the head effect to help sound localization. The localization algorithms in these two studies are mainly based on calculating distances between the cues in the received binaural sound and the cues in the database to decide the direction of the sound source. However, in such systems, the binaural cues extracted from received sound are correlated with sound source types in some extent. What’s more, the performance is also affected when there are multiple sources.

In our system, two omnidirectional microphones are placed at two end-points of a diameter of a rigid sphere. Although the structure is front-back symmetrical and neither pinna nor torso effects are introduced, which means that the sound source localization works only in the frontal azimuthal half-plane, we call it binaural localization, because it still satisfies in many applications. In our localization algorithm, a frequency-domain steered beamformer using both binaural cues is employed to make the system more robust[20]. Due to the introduction of the rigid sphere, the binaural cues vary with frequency, which requires to sub-band processing. The steered beamformer in each sub-band ranges from $-90^\circ$ to $90^\circ$ in azimuth in the horizontal plane (elevation $= 0^\circ$), and the final localization result is made by a joint judgement strategy. A multi-channel audio acquisition board is developed using DSP and USB to collect the audio data received by the microphones and then send the data to the computer, in which the localization program is executed and the final result is shown in the GUI. Both simulation and measurement results are given to show the performance of the system.

The rest of the paper is organized as follows. In Sec. 2, the scattering theory and the transfer function of a rigid sphere is introduced. In Sec. 3, our specialized filter-banks, sub-band beamformer, decision-making and the framework are presented in details. In Sec. 4, the system implementation including rigid sphere, microphone setup, the DSP board are introduced. In Sec. 5, both the experiment setup and evaluation results are presented. Some discussion is made in Sec. 6 and the following section is the conclusion.

2. BINAURAL CUES OF SPHERICAL SCATTERER

According to the scattering theory, the scattering problem in the frequency domain is reduced to solution of the Helmholtz equation for complex potential $\psi_s(r)$ given as

$$\nabla^2 \psi_s(r) + k^2 \psi_s(r) = 0$$

with the following impedance boundary conditions on the surface $S$ of the scatterer:

$$\left. \left( \frac{\partial \psi_s(r)}{\partial n} + i\sigma \psi_s(r) \right) \right|_S = 0$$

where $k = \omega/c$ is the wavenumber, $\sigma$ is the constant characterizing the impedance of the scatterer, and $i = \sqrt{-1}$. In the particular case of rigid scatterer ($\sigma = 0$) we have the Neumann boundary conditions,

$$\left. \frac{\partial \psi_s(r)}{\partial n} \right|_S = 0$$

Rabinowitz et al.[21] present the solution for the pressure on the surface of the rigid sphere due to a sinusoidal point source at any range $r$ greater than the sphere radius.
The binaural cues of a certain frequency $\omega$ can be extracted from the transfer function $H$. Here are the definitions for the binaural cues ITD (IPD) and IID:

$$ITD(r, \theta, \omega) = \frac{IPD(r, \theta, \omega)}{\omega} = \frac{\arctan(H_R/H_L)}{\omega}$$  \hspace{1cm} (9)

$$IID(r, \theta, \omega) = \frac{H_R^2}{H_L^2}$$  \hspace{1cm} (10)

When computing the IPD, unwrapping should be used to correct the phase by adding multiples of $\pm 2\pi$ to smooth the ITD of a series of frequencies. Fig. 1 shows comparison of IID cue in free sound field and in the rigid sphere scattered sound field. It can be easily seen that the IID cue becomes much more notable when a scatterer is introduced. Due to the small effect of $r$ on $H$ or binaural cues, especially when the source is in the far field, the effect of the argument $r$ is omitted in the system algorithm design.

### 3. SOUND LOCALIZATION BASED ON SUB-BAND BEAMFORMER

#### 3.1. Beamformer with Both Binaural Cues

The output $y(t)$ of a basic $N$-microphone delay-and-sum beamformer is defined as:

$$y(t) = \sum_{n=1}^{N} x_n(t - \tau_n)$$  \hspace{1cm} (11)

And this can be realized in the frequency domain by only calculating the cross-correlation function[20]. However, as said in Sec. 1, only the time cue is used in the beamformer while the intensity cue is neglected although it becomes notable when the scatterer is introduced. For the binaural system, according to the definition of the binaural cues in equation (9) and (10), the beamformer with both binaural cues at a certain frequency $\omega$ is modified as equation (12) to maximize the output $Y$.

$$Y(\omega) = X_L(\omega) + \frac{1}{\sqrt{IID(\theta, \omega)}} X_R(\omega)e^{i\omega ITD(\theta, \omega)}$$  \hspace{1cm} (12)

Then, the beamformer output energy changes to

$$E_Y = X_L^2 + \frac{1}{IID} X_R^2 + \frac{2}{\sqrt{IID}} X_L X_R e^{i\omega ITD}$$  \hspace{1cm} (13)

which is different from the basic beamformer in [20] because the IID is not a constant as direction $\theta$ changes.
The beamformer output energy computed in the frequency domain not only relates to the cross-correlation function but also relates to the $X_k^2 + X_R^2 / \text{IID}$ part, which is not a constant. Any further improvement on this IID-related beamformer, such as spectral weighting[20], should consider all parts of the beamformer in equation (13).

3.2 Specialized Filterbanks
As shown in Sec. 2, both the binaural cues vary with frequencies, so the beamformer in (12) is supposed to be done in each frequency component. However, it is not realistic to do so due to large computation. Here, sub-band processing is introduced to reduce the computation complexity. In each sub-band, it is assumed that all the frequency components share the same binaural cues as those of the center-frequency component.

The Gammatone filterbanks[22] are the most frequently used filterbanks in auditory signal processing due to the good simulation of the cochlea sub-band processing. However, there are much overlapping among bands in gammatone filterbanks, which makes the assumption mentioned above unreasonable and brings down the performance of the beamformer directivity.

Here, a kind of specialized filterbanks is designed as follows. The center frequencies of the filterbanks are the same as those of the gammatone filterbanks due to their distribution on the log-frequency domain. The bandwidth of each sub-band is decided by center frequencies of its two neighboring sub-bands, which restricts unnecessary overlapping among sub-bands so that the beamformer based on the assumption mentioned above performs better. Each sub-band filter can be realized by the basic FIR filter. The magnitude-frequency response of this type of filterbanks is shown as Fig. 2.

3.3 Multi-band Joint Judgement
Assume that the input signal are filtered into $M$ sub-bands by the filterbanks. In each sub-band, a beamformer is steered in the possible range to search the peaks of its output energy. It is well known that the peak width is the function of frequency. Specifically, there are just a few wide peaks in low frequencies, while many narrow peaks exist in the high frequencies. It is suggested in [5] to use the “coarse-fine search” to improve the algorithm efficiency. But this strategy requires that results of the low frequency beamformers be as correct as possible, or, the search range of high frequency beamformers may be misled by previous results. In other words, the decision-making method in the “coarse-fine search” should vary with the sound source spectrum, which is not robust to sound source type. In all bands, the same step, which is also the spatial resolution of the system, is taken in the steered beamformer. Obviously, this method brings more computation, but it still can work in real time due to our $-90^\circ$ to $90^\circ$ task. Steered beamformers in various sub-bands export their own results independently, and then the final localization result is obtained by the multi-band
joint judgement strategy. In this joint judgement strategy, the \( i \)-th beamformer output energy \( E_i(\theta) \) is normalized to be converted to the probability \( P_i(\theta) \) as

\[
P_i(\theta) = \frac{E_i(\theta)}{E_0B(\theta)}
\]

Here, the \( E_0 \) is the energy threshold to ensure that the probability \( P_i(\theta) \) be not greater than 1. The \( B(\theta) \) is the directional energy compensation factor. As shown in Fig. 3, when sweeping in different azimuths using white noise, the normalized output energy \( B(\theta) \) varies with the direction. In order to ensure the equality of all directions, this gain brought by the beamformer itself should be removed as shown in equation (14). Then the joint probability \( P(\theta) \) is calculated as equation (15) to get the final result.

\[
P(\theta) = \prod_{i=1}^{M} P_i(\theta)
\]

In order to find peaks in \( P(\theta) \) more easily and accurately, \( P(\theta) \) is smoothed by convolving with a Hanning window, the length of which is empirically chosen. The smoothing makes the “real peaks” more distinct and especially, removes those “fake candidates” near the “real ones”. And the final source localization result \( \hat{\theta} \) is obtained by

\[
\hat{\theta} = \arg\max_{\theta} P(\theta)
\]

3.4. Algorithm Framework

The algorithm framework is given in Fig. 4. First, the standard ITDs and IIDs of different directions and frequencies are calculated off-line using the the rigid sphere transfer function \( H(\theta, \omega) \) in equation (8). Then the input data in the buffer are filtered to sub-bands using the specialized filterbanks. Sub-band beamformers are steered from \(-90^\circ\) to \(90^\circ\) in azimuth respectively with beamformer output energy data calculated and the joint judgement and the smoothing strategies help get the final localization result.

4. SYSTEM IMPLEMENTATION

In this system, a rigid plastic sphere with a diameter of 18.2 cm is used as the sphere scatterer and two omnidirectional microphones are placed at two end-points of a diameter, on the surface of the rigid sphere, as shown in Fig. 5. The whole sphere is stabilized on a bracket and stays away from the ground in a certain distance to eliminate the effect of other scatters as much as possible.

The sound collected by the two microphones are transferred to a multi-channel audio signal acquisition board. Note that there are six audio input channels on the board, any two of which can be used here while the others are for extensive multi-channel use. This board, instead of the sound card in the computer, is employed here as sound cards in computer often could not qualify in the aspect of signal-to-noise ratio.

Fig. 6 shows the frame of the acquisition board. Cored by TI TMS320VC5509 chip, this board is composed of A/D converter, CPLD, FLASH and power management modules. The reason to choose DSP chip here lies in the consideration of embedding signal processing algorithm in the DSP in future. The two channel sound captured by the two microphones is performed by A/D conversion through the AD73360 chip with a sampling rate of 16 kHz to the Multi-channel Bus Serial Port Interface(McBSP) of the DSP using the SPI protocol. The Direct Memory Access(DMA) mechanism is then introduced to transfer the digitalized signal to the computer for further signal processing through the USB interface. The CPLD part is a control unit responsible for setting the mode of bootloader, the timer input and the working mode of DSP. The FLASH part is used to store programs, which are transferred to the memory of DSP when boot-strap or reset happens.
Fig. 5: The rigid spherical scatterer and the microphones.

Fig. 6: Frame of the multi-channel audio signal acquisition board.

The inter-channel differences of the whole signal acquisition module, which includes the microphones and the board, are pre-measured in an anechoic chamber and then compensated in the computer before any further signal processing. The core localization procedure is then carried out in the computer and the final result is shown in a GUI.

5. EXPERIMENTS

5.1. Simulation Experiments

To demonstrate the efficiency of the proposed algorithm, simulation and measurement experiments are performed under various conditions. In the simulation experiments, the synthesized input signals are directly sent to the core localization procedure without microphones or DSP board. The azimuth of the sound source, which is synthesized through a single channel sound convolved with the binaural rigid sphere transfer functions, ranges from 0° to 90°. There are 4 types of the original single channel sound: vowel [a:], claps, a speech sentence by a male speaker and white noise. Also sound sources in noisy environment are synthesized by adding two uncorrelated background noise (white noise) to the two channel inputs without spatial information at SNRs of 0dB, 5dB, 10dB, 15dB and 20dB.

Table 1 and Table 2 show that the time-average localization errors vary with different target azimuth and SNRs respectively, for 4 types of sound sources. Fig. 7 and Fig. 8 show that the localization standard variances for 4 types of sound sources with 15dB SNR and 0dB SNR, respectively. Fig. 9 shows that localization standard variances under all types of SNRs. It can be seen easily that the localization performance is good when the target azimuth is no greater than 60°, while the peak width of the beamformer becomes larger and larger when the azimuth increases. The speech source generally perform worse than other types because the energy of speech varies with time significantly, while the white noise has the best performance due to its stationary energy compared to other sources. Seen from Fig. 9, the system is robust to the background noise.

5.2. Measurement Experiments

Measurement experiments are carried out in an ordinary office room without obstacles between the sound source and the spherical scatterer. The source is placed about 1.4m far away from the center of the sphere. Actual azimuths of the source are 0°, 20°, 40° and 60°. The male
speech, claps and white noise used in the simulation experiments are broadcasted in a small loudspeaker to act the source. Fig. 10 shows the localization results varying with time. It can be seen that the white noise source performs best while the speech performs worst, which is consistent with the simulation results.

6. DISCUSSION
The proposed localization algorithm is built up with some ideas inspired by the human audition, in which the cochlea-like filterbank and the binaural cues are employed. It is certain that the goal of the algorithm is to localize the sound source instead of the understanding or characterization of human audition, but we do hope that the introduction of those human audition inspired ideas can make some help to achieve the goal.

As mentioned before, due to the limitation of the sym-
metrical structure, this binaural system only works in the frontal azimuthal half-plane. It is easy to solve the front-back confusion problem using two microphones, if they are not placed at two end-points of a diameter of the rigid sphere to generate “unsymmetrical structure”, but the performance will decrease at the “broad side”, on which the angle between the two microphones is larger. And if 3-D localization is needed, the number of microphones can be increased, which is one of our future work.

From the experimental results, it’s not hard to see that the localization error increases from 65° to 80° under all circumstances. The reason for this phenomenon includes two aspects. First, the width of the beamformer increases as the target azimuth gets close to the diameter, on which the two microphones are placed. Second, due to the front-back symmetry, even if the steering range is from −90° to 90°, there exists a “virtual” beamformer on the back side. When the target azimuth gets close to the diameter, there is some overlap between the desired beamformer and the “virtual” one, which makes the maximum energy output locates around 90°, so the localization error is relatively larger from 65° to 80°. Unfortunately, the problem is currently unavoidable in such symmetrical settings, while it can be solved by simply using unsymmetrical structures.

In the measurement experiment results, as shown in Fig. 10, the localization performance for white noise is better than those for other types of sound sources, because white noise is the most time-stationary sound sources among the three types and currently, there is no time-tracking strategy (such as Kalman filter or particle filter) employed in this system. The time tracking part is also one of our future work.

7. CONCLUSION
A binaural sound source localization system in the frontal azimuthal half-plane based on sub-band steered beamformer with spherical scatterer is introduced and realized. Both the ITD and IID cues are used in the beamformer. The rigid sphere transfer functions, from which the binaural cues are extracted, are used instead of HRTFs and calculated using the scatterer theory. Specialized filterbanks are introduced for sub-band beamformer and the multi-band joint judgement strategy is employed to make the decision-making more robust to various source types. Both the simulation and measurement experiments show good performance of the system. In the future, the localization and tracking task of multi-sources with movement based on this framework is going to be investigated and the embedding of the localization algorithm to the DSP board is also in consideration. Other aspects, such as the various scatterer models and beamformers are going to be tested to improve the performance.

8. ACKNOWLEDGEMENT
This work was supported by China Natural Science Foundation for Young Scientists (60305004); the Key Project of China Natural Science Foundation (60435010, 60535030), the Project of New Century Excellent Talents in Universities.

9. REFERENCES


