Robotica (2010) volume 28, pp. 1013–1020. © Cambridge University Press 2010
doi:10.1017/S0263574709990865

Azimuthal source localization using interaural coherence in a robotic dog: modeling and application

Rong Liu†,‡,* and Yongxuan Wang‡

† *State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150080, P.R. China*
‡*School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, P.R. China*

(Received in Final Form: December 8, 2009. First published online: January 15, 2010)

SUMMARY

In nature, sounds from multiple sources, as well as reflections from the surfaces of the physical surroundings, arrive concurrently from different directions at the ears of a listener. Despite the fact that all of these waveforms sum at the eardrums, humans with normal hearing can effortlessly segregate interesting sounds from echoes and other sources of background noises. This paper presents a two-microphone technique for localization of sound sources to effectively guide robotic navigation. Its fundamental structure is adopted from a binaural signal-processing scheme employed in biological systems for the localization of sources using interaural time differences (ITDs). The two input signals are analyzed for coincidences along left/right-channel delay-line pairs. The coincidence time instants are presented as a function of the interaural coherence (IC). Specifically, we build a sphere head model for the selected robot and apply the mechanism of binaural cues selection observed in mammalian hearing system to mitigate the effects of sound echoes. The sound source is found by determining the azimuth at which the maximum of probability density function (PDF) of ITD cues occurs. This eliminates the localization artifacts found during tests. The experimental results of a systematic evaluation demonstrate the superior performance of the proposed method.

KEYWORDS: Sound localization; Interaural coherence; Cross-correlation; Head model.

1. Introduction

Acoustic stimulus is one of the most important sensory inputs for humans and animals to interact with their environments.[1] Nevertheless, the usage of auditory information in mobile robot navigation has not yet been fully studied, compared with the well-investigated vision-based navigation techniques.[3,4] This may be mainly attributed to the lower spatial resolution of the acoustic information. More recently, due to the increased complexities of the environments where the robots are deployed, the inclusion of auditory information in robotic navigation system receives more attention. The abilities of the audition-based navigation techniques become widely recognized in compensating the narrow visual field as well as cooperating with the visual systems for a more efficient navigation.[5,6]

Acoustic source localization is the process of determining the spatial position of an emitting source using the received acoustic signals.[2] Most artificial sound localization systems involve microphone arrays acting as beamformers.[7] Such systems are usually computationally intensive in the sense that they have to process a multitude of signals. Other biomimetic robotic systems comprise microphone pairs mounted on the robot head, which is the case considered in this paper. One advantage of such construction is computational costs saving.[1]

In biological systems, binaural sound source localization relies on two major cues—interaural level difference (ILD) and interaural time difference (ITD). ILDs are caused by the acoustic shadow of the head, attenuating the sound arriving at the eardrum which is farthest from the source. ITDs arise from the differences in conduction time a sound wave takes to reach the two ear drums.[8] Biologically inspired sound localization systems have implemented either one of these cues,[9,17] or both.[16]

There are several ways to find the location of a sound source with binaural cues. A common paradigm to approximately find the location of a sound source employs a dual-microphone array and estimates time differences of arrival (TDOA) between the microphones in the array.[22,23] Generally, it is assumed that the difference in the two channels of a dual-microphone array is limited to a small time delay (or linear phase in frequency domain). Thus, methods that search for extrema in cross-correlation waveforms are commonly used.[17] However, this kind of approaches are based on the assumptions that the sound waves propagate along a single path from the sources to the microphones and that the microphone responses of the two channels for the given source sources are approximately the same. In order for these conditions to be satisfied, the microphones need to be identical, co-aligned and close to each other with respect to the acoustic source. The time delay assumption fails in the case of our robot's (SONY robotic dog, Aibo ERS-7) head: the microphones are not identical and additionally, the Aibo's head and body will filter the response in a complicated way. Alternative localization approach involves head related transfer function (HRTF).[10–13] HRTF depends on the shape of head and changes as environments change. Moreover, the cost of measuring HRTF for each robot is expensive and only HRTFs for discrete directions are available. The active

* Corresponding author. E-mail: rliu@dlut.edu.cn

https://doi.org/10.1017/S0263574709990865 Published online by Cambridge University Press



Fig. 1. The view of Aibo ERS-7's head. Each microphone mounted on the head with a plastic earlobe.

audition methods proposed by Nakadai[14,15] and Berglund[16] explore interaural phase differences (IPDs) in a number of distinct frequency sub-bands. The problem of these methods is phase ambiguity and computation complexity.

With these difficulties in mind and inspired by the mechanisms of human auditory system to explain various aspects of source localization in complex acoustic situations, we develop a simple but effective algorithm to extract the directional acoustic information for the Aibo. Our approach is composed of three stages. First, we build a head model of the selected robot. Then, we perform signal processing based on the generalized cross-correlation implementation. Afterward, rather than directly jumping into peak-finding we employ a probability method by casting the azimuth estimation as a cues selection problem. The improvements in performance achieved by combining the three stages together are validated by extensive experimental tests.

The rest of this paper is organized as follows. In Section 2, a brief description of the auditory system setup is given and the microphone mismatch problem is addressed. The algorithm for sound source localization is developed in Section 3. The experimental results and discussions on the results are presented in Section 4. A brief conclusion and future work is then followed.

2. Auditory System

An AIBO (meaning "companion" or "pal" in Japanese), shown in Fig. 1, is used as the server for monitoring information transmission as well as a companion of the patient in our healthcare system. In this section, we briefly review the product specifications regarding the Aibo's auditory system configuration, which is the basis for the robot's head model development and the sound localization algorithm introduced in the following sections. Biologically inspired approaches restrict themselves to two inputs, equivalent to the two ears as shown in Fig. 1. The two microphones are mounted antipodally on the robot head with two plastic earlobes for decoration. The baseline

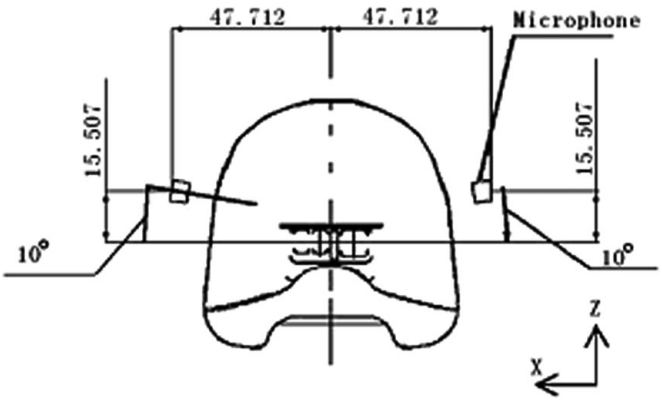


Fig. 2. Head dimensions of the Aibo ERS-7 in front view.

distance between the microphones is 95.424 mm (see Fig. 2). Digital stereo samples are streamed from the microphones at 16 bits/channel, with a sampling rate of 16 kHz. A binaural system can effectively compute interaural differences in time and intensity, thereby making it possible to accurately determine the directions of sound source. ILD is normally interpreted as an energy-like (time-integrated) measure of the stimulus level. It can be used as an indicator of three possible sound source directions: left, right, or center[14] to eliminate the ambiguity in the time delay estimation. The usefulness of the ILD in acoustic source localization relies on whether the two audio channels are identical. Therefore, in the use of dual microphones to achieve the desired directional characteristics, a major consideration is to ensure that both microphones are identical in their sensitivity and phase characteristics. Unfortunately, all electroacoustic transducers suffer from sensitivity drift over time as a result of aging, and excessive temperature and humidity may accelerate the drift. Additionally, significant variations may exist among microphones because of imperfect manufacturing, so it is often impossible to predict how a microphone performs over time. Our tests have confirmed that there is a mismatch between the microphones pair under different frequencies. The channel mismatch may introduce gain and phase errors, which degrades the performance of ILD-based techniques. Hence, ILD information is not reliable enough to be included in the sound source direction estimation process.

3. Sound Localization System

The biological systems feature some of physical, physiological, and psychological processing stages for sound source direction discrimination. Torso and external ears play a significant role in sound localization for biological systems, especially for elevation determination.[8] Due to the specific configuration of this robot auditory system, localization of stationary sources in azimuth is possible only up to a front-back ambiguity. Therefore, the study here only involves localization in azimuth with a range of 180°.

Different models have been used to explain various aspects of binaural perception. In the following subsections, we propose a coincidence mechanism based on time-specific manner. This approach is very straightforward: only ITD cues occurring at time instants when they represent the direction

https://doi.org/10.1017/S0263574709990865 Published online by Cambridge University Press

of one of the sources are selected, while other cues are ignored. The proposed cue selection method is implemented in the framework of a model that considers the physical and physiological stages of the robot.

3.1. Aibo's head model

For the direction estimation, a model is required to predict the distance difference that the sound source on a specific azimuth must travel to reach each ear of the robot. Various models have been developed for different applications including headless model,[17,18] Lord Rayleigh's model,[19] Epipolar geometry.[20] The headless models assume an empty space between the microphones leading to an underestimation of the azimuth. The Lord Rayleigh's model assumes that the left and right signals arrive at the head in two parallel lines. However, this is only the case if the sound source is of the exact same size as the head or if the sound source is very far away. Therefore, the sphere model is considered in our application which is the situation most commonly encountered. The sound source is represented as a point source S. The robot head is modeled as a sphere with the radius R_m measured to be 47.712 mm (see Fig. 2). The path to the left microphone (SA) is composed of two segments while the path to the right one (SB) is a straight line. The sound wave travels to a point of tangency M around the sphere to the observation point A.

3.2. ITD estimation

The cross-correlation of sound waves from dual-microphones is calculated to estimate the ITD. ITD is the primary measure in this study to determine the coherence between the two signals. The correlation function, denoted as $R(\tau)$, is defined by

$$R(\tau) = \sum_t s_r(t+\tau)s_l(t), \tag{1}$$

where s_r and s_l are the signals recorded in the right and left microphones, respectively, and t ranges over the time window of the input signals. The interval of interest over which the correlation index τ is allowed to vary for time delay estimation is normally much smaller than the duration of the input signals. The value of τ that maximizes $R(\tau)$ corresponds to the time delay estimate. This computed delay is either negative (less than the mid-point) or positive (higher than the mid-point).

Let $s_1[k]$ and $s_2[k]$ denote the sequences of N digitized samples of two channels. The cross-correlation of the two sequences is defined by

$$Cs_1s_2 = \frac{1}{N}\sum_{n=0}^{N-1} s_1[n]s_2[n+m]. \tag{2}$$

3.3. Azimuth estimation

The spatial orientation of the sound source S is represented by the spherical coordinate. As seen in Fig. 3, θ is the azimuth, φ is the elevation, and d is the distance of sound source to the center of the sphere, respectively. Converting

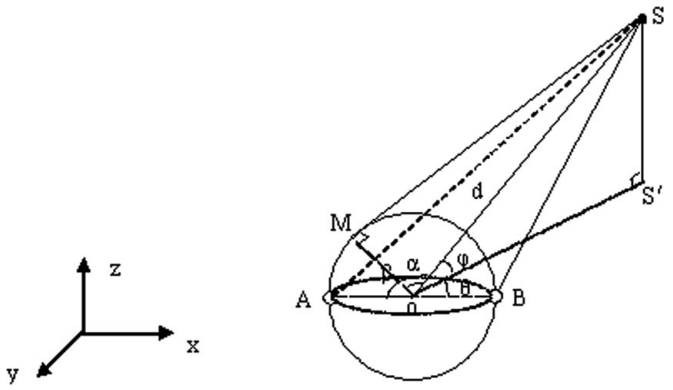


Fig. 3. The sphere model of Aibo ERS-7's head.

the spherical coordinate into the rectangular coordinate, S is given as $(d\cos\varphi\cos\theta, d\cos\varphi\sin\theta, d\sin\varphi)$. The spatial position of left microphone A and right microphone B are $(-R_m, 0, 0)$ and $(R_m, 0, 0)$, respectively. The distance of the straight line distance SB from the source to the observation point B is

$$\begin{aligned} L_{SB} &= \sqrt{(d\cos\varphi\cos\theta - R_m)^2 + (d\cos\varphi\sin\theta)^2 + (d\sin\varphi)^2}, \\ &= \sqrt{d^2 + R_m^2 - 2dR_m\cos\varphi\cos\theta}. \end{aligned} \tag{3}$$

Similarly, the distance between the source and the tangent point M is computed by

$$\begin{aligned} L_{SM} &= \sqrt{(d\cos\varphi\cos\theta + R_m)^2 + (d\cos\varphi\sin\theta)^2 + (d\sin\varphi)^2}, \\ &= \sqrt{d^2 + R_m^2 + 2dR_m\cos\varphi\cos\theta}. \end{aligned} \tag{4}$$

According to cosine theorem of angle, we obtain the angle β contained by Ao and So, and the angle α contained by Mo and So through

$$\cos\beta = \frac{d^2 + R^2 - L_{SA}}{2dR_m} = -\cos\varphi\cos\theta, \tag{5}$$

$$\cos\alpha = \frac{R_m}{d}. \tag{6}$$

If the speed of propagation is assumed to be c both in air and around the surface, the time difference Δt between the time that the wave reaches the two microphones is given by

$$\begin{aligned} \Delta t(\theta,\varphi,\rho) = \frac{R_m}{c}\Bigg[&\left(\frac{\pi}{2} - \theta\right) - \arccos\frac{1}{\rho} + \sqrt{\rho^2 - 1} \\ &- \sqrt{\rho^2 + 1 - 2\rho\cos\varphi\cos\theta}\Bigg], \end{aligned} \tag{7}$$

where $\rho = d/R_m$, $\rho \geq 1$.

Therefore, the ITD is a function of azimuth, elevation as well as distance. For computational simplicity in converting the Δt to the azimuth, we assume that the distance of the

https://doi.org/10.1017/S0263574709990865 Published online by Cambridge University Press

sound source to the head is infinity and elevation angle φ to be zero, which leaves us with

$$ITD(\theta) = \frac{fR_m}{c}\left[\left(\frac{\pi}{2} - \theta\right) + \cos\theta\right]. \quad (8)$$

where c is the speed of sound which is approximately 345 m/s in indoor environment, ITD is the time delay in sample units, and f represents the sampling frequency.

3.4. Interpolation

In practice, the maximizer of the correlation function in Eq. (2) may not be exactly located on sampling points. In some cases, it may fall between two consecutive sampling points. To break the fundamental limit on the ITD estimation accuracy imposed by the sampling process, we apply quadratic interpolation to the three correlation values ($N-1$, N, $N+1$) centered at the maximum of the correlation N. If the discrete-time signal obtained using a sampling rate of f_s is interpolated to an oversampling rate f_s' with a ratio L, then there are $L-1$ new data points that need to be computed between any two sample points.

The cross-correlation of the two sequences after interpolation is defined by

$$Cs_1s_2[n, m] = \frac{1}{NL}\sum_{n=0}^{NL-|m|-1} s_1[n]s_2[n+m], \quad (9)$$

where m ranges from $-m_{\max}$ to $m_{\max}$, and $m_{\max}$ represents the largest lag. Define m^* by

$$m^* = \arg\max_m C_{s1s2}[n, m], \quad (10)$$

where arg max means argument of maximization.

This yields the time delay to be

$$\Delta t = \frac{m^*}{Lf}. \quad (11)$$

Correlation index m is calculated over a small range corresponding to the expected rang of delays for the baseline. This means a propagation delay of six samples when L is set to be 1.

3.5. Selection of binaural cues

In indoor environments, sounds from source as well as reflections from the surfaces of the physical surroundings arrive concurrently from different directions to the ears of Aibo. As a result, accurate sound localization becomes more difficult, compared with the case where only line-of-sight (LOS) propagation exists. Even worse, the combined effects of multiple reflective sound paths can be as loud as or louder than the sound from the LOS path. Therefore, the auditory system needs to be able to suppress the effect of the reflections to guarantee a satisfactory performance.

Acoustic signals from the source arrive at the microphones before the superposition of echoes should be highly coherent. Faller and Merrimaa[21] have illustrated this when developing a model to mimic the human auditory mechanism for sound localization in complex listening situations. With these observations in mind, we shall only select the ITD cues that are highly coherent to represent the direction which makes the algorithm more robust to the negative influences of the reverberation and echoes.

Interaural coherence (IC) is used as an indicator for the time instants when the cues are highly coherent. More specifically, by selecting ITD cues to coincide with IC cues larger than a certain threshold, a subset of ITDs similar to the corresponding cues of the source is obtained. Note that compared with the models implemented for human auditory system, our method does not require realizing the auditory periphery signal processing.

IC is defined as the maximum coefficient of each cross-correlation function as

$$c_{12}(n) = \max_m c_{s1s2}[n, m]. \quad (12)$$

IC has a range of [0, 1] in theory, and 1 indicates perfect coherence of two signals. At each time instance n, these two signal properties are denoted as cue pairs $\{\tau(n), c_{12}(n)\}$. Given all $\tau(n)$, we only consider the subset of cues whose ICs are larger than a certain threshold c_0. Mathematically, the subset is given as

$$\{\tau(n)\,|c_{12}(n) > c_0\}. \quad (13)$$

When the "first wave front" arrives at the ears of the robot, the evoked interaural cues are similar to the free field LOS cues, i.e., $c12(n) \approx 1$. Later on, the first reflection wave reaches the robot, which would superpose on the LOS signal. In this case, the IC coefficient would deviate significantly from 1.

It can be expected that with a large threshold c_0, the selected cues are more likely to be free field cues. However, in practical implementation, we tend to select the minimum c_0 that gives sufficiently accurate interaural cues. This strategy could lead to a larger number of selected cues and consequently a larger portion of the input signals being involved in the localization process.

Given the statistics of these binaural parameters, the sound source direction can be estimated by the cues selection-based classification. By grouping the time differences for each instant into a feature vector

$$(\vec{x}) = (\tau(1), \tau(2), \tau(3), \cdots). \quad (14)$$

The one dimensional *probability density function* (PDF) $P(\tau)$ and the PDF of the feature vector $\vec{x}\,P(\vec{x})$ are estimated by computing histograms of ITD cues. The ITD with highest probability is used to estimate the direction of the sound source.

4. Experimental Results and Discussions

The performance of our localization method is evaluated with real-world recorded sounds. In these experiments, Aibo and sound source (a loudspeaker) were located in our laboratory environment, a reverberant environment with various background noise sources. No special treatment was made to the room. The sound sources were positioned in

https://doi.org/10.1017/S0263574709990865 Published online by Cambridge University Press

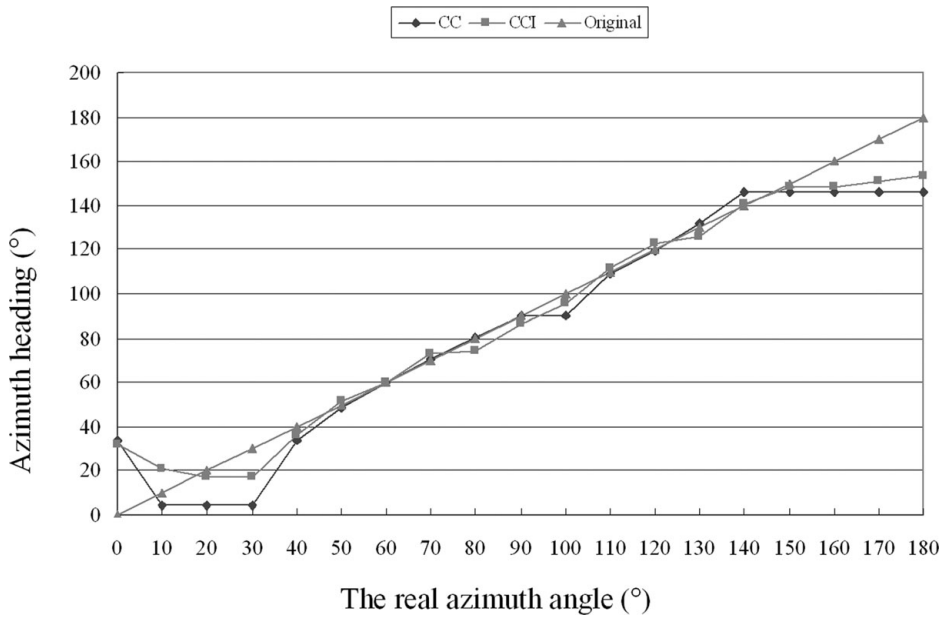


Fig. 4. The measured direction with respect to the real one.

the front hemisphere with respect to Aibo. Two types of sounds were applied in the experiments. One is a pure tone with 1 kHz and the other is a harmonic sound including multiple frequency components from 50 Hz to 4 kHz. For simplifying the following presentation, these two sounds are termed shortly as 'simple' and 'music', respectively, in the rest of this section. The duration of a single play for each type of sound is 2 s and 5 s. They were played repeatedly twice.

4.1. Aibo head model test

In this test, we check the rationality to find the position of the sound event based on the sphere head model. The sound sources were placed at positions over the full hemispheric range of directions from 0° (right lateral side of the robot) to 180° (left lateral side of the robot), at 45° increment. The cross-correlation is done with the whole data length. The result for the test shows that this model is able to yield satisfactory sound direction estimates. There exist some larger deviations at two sideward directions.

According to Eq. (8), the discretization of the corresponding azimuth with respect to the delay samples is not linear. The azimuth increments are smaller at front directions and increase to near 30° at both lateral directions. This resolution could be improved by other techniques, for instance, the interpolation scheme in our approach. The effect of the interpolation technique is shown in Fig. 4. A set of data comprising sound waves emitting from 0° to 180° at intervals of 10° were used for this test. Considering the number of possible lags and the tested angles, interpolation rate is set to be 2. It shows the measured sound location for each direction with normal correlation, correlation with interpolation ($L = 2$) and the theoretical curve. Legends of these three methods are denoted as CC, CCI, and Original, respectively. From this figure, we observe that the measured angles share the same trends as the theoretical lines. The accuracy is good in front of the robot, and gets worse at the lateral sides. Moreover, the method with interpolation is better than normal one at enhanced resolution, for example, at position when the direction angle is 20°. Overall, we can see our model is effective in azimuth estimation.

4.2. Sound source direction estimation with ITDs selection

Cross-correlation is sensitive to signal length. That is, the longer the signal length, the more time is spent on it. Therefore, long signals should be avoided in real-time applications. We only use 120 samples (0.0075 s) to perform cross-correlation.

In this test, the proposed approach is applied to localize the musical stimuli. The cues selection threshold c_0 is set to be 0.85. Two kinds of plots are used to illustrate the procedure. First, the instantaneous ITD and ILD values are plotted as a function of time with the selected cues represented as black solid lines. In addition, the effect of the cue selection is visualized by plotting PDFs of the selected ITD and ILD cues. The PDFs are estimated by computing histograms of ITD and ILD cues in time span of 0.5 s. ILD is plotted here to show that it is unreliable for sound localization. Figure 5 shows the IC, ITD, and ILD as a function of time when the musical sound source is located at an azimuth angle of 30°. Superposition of signals from different directions results in fluctuating time and level differences. Figure 6 shows PDFs of ITD and ILD in the case of without cues selection and with cues selection. When using all the cues available, the localization algorithm yields two distinct peaks for the likelihood scores of this sound. The maximum of the PDFs does not appear at the correct location, which results in the azimuthal angle to be 175°. One possible reason is that although cross-correlation provides excellent time delay estimation for both broadband signals and narrowband stimuli in the low-frequency ranges, for periodic waveforms it may present ambiguous peaks at multiples of fundamental frequencies. Furthermore, sound reflections and reverberation often occurred in indoor environments

https://doi.org/10.1017/S0263574709990865 Published online by Cambridge University Press

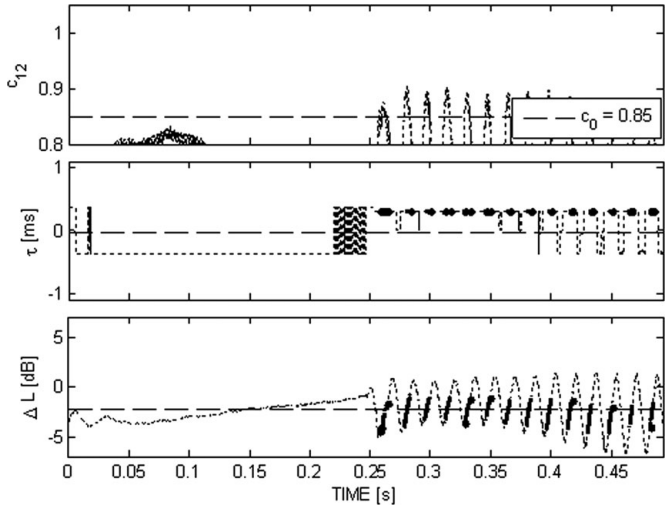


Fig. 5. IC, ITD, and ILD as a function of time for the musical sound source at 30° azimuth. The cue selection thresholds (top) and the cues of sources (middle and bottom) are indicated with dotted line. Selected cues are marked with bold solid lines.

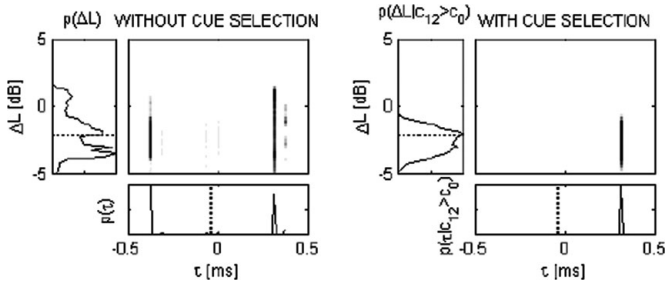


Fig. 6. PDFs of ITDs and ILDs for the musical sound source at 30° azimuth.

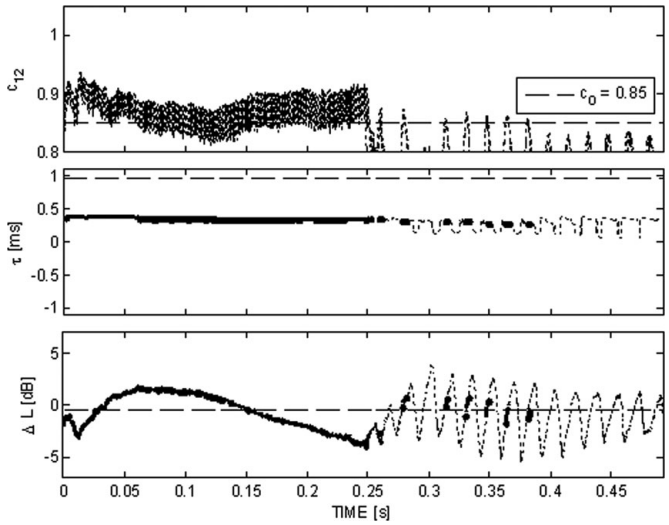


Fig. 7. IC, ITD, and ILD as a function of time for the musical sound source at 20° azimuth. The cue selection thresholds (top) and the cues of sources (middle and bottom) are indicated with dotted line. Selected cues are marked with bold solid lines.

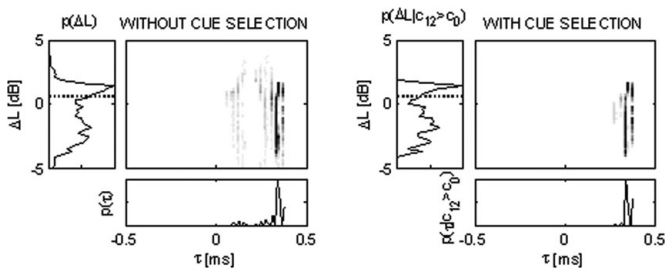


Fig. 8. PDFs of ITDs and ILDs for the musical sound source at 20° azimuth.

may corrupt the delay estimation. With the cues selection method, the peak of ITD is appropriately selected. The peak of PDFs corresponds to the angle of 33° which approaches to the true angle. We also carried out studies on the performance of the cue selection method in conjunction with interpolation technique to improve localization resolution. The algorithm was assessed with a musical sound at 20°, which cannot be located with normal cross-correlation method. The interpolation rate is set to 2.

Figure 7 shows the IC, ILD, ITD as a function of time for this sound source and the PDFs without and with cue selection cases are shown in Fig. 8. The results for both cases are the same. The maximum PDF located at 0.35 ms corresponds to the angle 21°. Thus, the computed direction of auditory event coincides with the directions of the corresponding physical source. Besides the above-mentioned two examples, we have tested the algorithm with other recorded data. The peaks of ILDs are in line with the associated directions. From these tests, we found that the ILDs sometimes indicate the wrong directions. Therefore, ILDs are less reliable and ITDs become more critical in the localization methods.

4.3. Discussions

Actually, the proposed approach may not use a hard IC threshold to select or discard binaural cues. Instead of pure selection, similar processing could be implemented as an IC based weighting of ITD cues with a slightly smoother transition. The performance of proposed approach was assessed as a function of c_0 for a sound source with the critical bands center frequencies to be 200, 500, and 1500 Hz, respectively. Figure 9 show the ITD bias and standard deviation, as well as the fraction of signal power corresponding to the selected cues as a function of c_0. The optimal thresholds for different critical bands are different. As expected, due to the neural transduction IC has a smaller rang at the 1.5 kHz critical band than at the 500 Hz critical band. Consequently, a larger c_0 is required.

It is assumed that the auditory system adapts c_0 for each specific listening situation, i.e., for each scenario with a constant number of active sources at specific locations in a constant acoustical environment. Since the listening situations do not usually change very quickly, it is assumed that c_0 is adapted relatively slowly in time. Therefore, all simulations reported in this paper consider only one specific listening situation at a time. As a result, a single constant c_0 is used for each simulation. For all the previous simulations, a suitable c_0 should be chose as a compromise between similarity of the cues to free-field cues and how frequently cues are selected.

In terms of computation, our Matlab implementation running on a computer with a 2.6 GHz Intel Core 2 Duo

https://doi.org/10.1017/S0263574709990865 Published online by Cambridge University Press

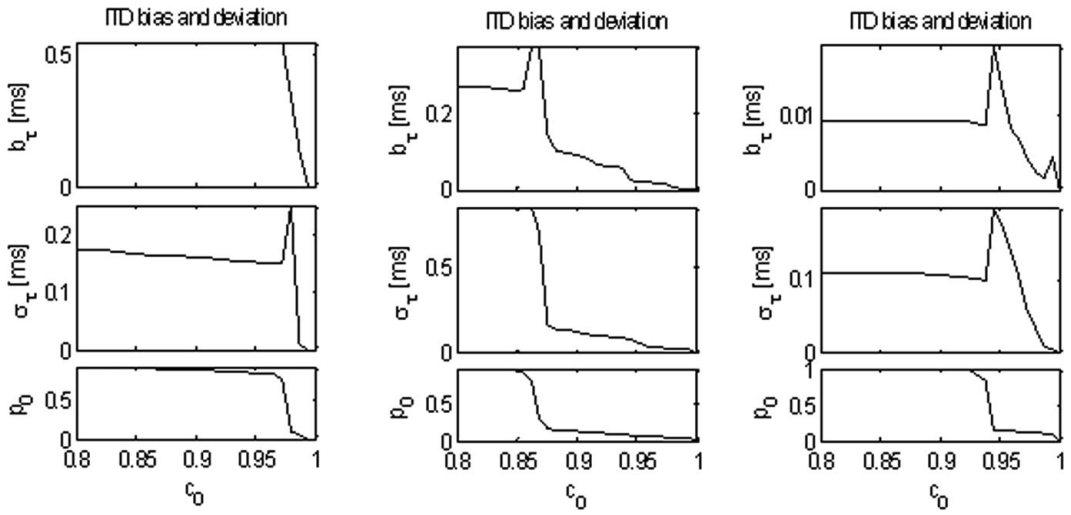


Fig. 9. ITD bais, standard deviation, and relative power of the selected signal portions as a function of the cue selection threshold c_0 for 200, 500, and 1500 Hz critical bands. (a) The 200 Hz critical band; (b) the 500 Hz critical band; (c) the 1500 Hz critical band.

processor and 2 G of RAM. Computation time is dominated by the cues selection algorithm. The computation time of our algorithm is approximately 2.3 s. Compared with other algorithms, it is reasonably efficient.

It is difficult to compare the performance of our algorithm with the performance of the other methods, because different robots or test environments were used. Although comparisons are difficult, we have the impression that our system, despite its simplicity, does not perform worse in azimuth estimation than other methods using complex computation. In refs. [14, 15], a frequency-domain algorithm for the sound localization subsystem of the humanoid torso special interest group (SIG) was used. The method performs ITD extraction by directly computing the phase differences between the left and right channels from fast fourier transform (FFT) frequency peaks. Additionally, ILDs were include in the azimuth estimation. Although SIG group reported an accuracy within $\pm 5^\circ$ from 0° to 30° and deteriorated for more lateral positions. However, Berglund[16] reported that he had implemented the SIG algorithm on an Aibo robot and only achieved a somewhat lower accuracy of about 10° for the frontal directions. With high singal-to-noise and direct-to-reverberant rations, the precision is 5° for the method proposed in ref. [16]. ITDs were computed by the zero crossings of the waveforms from microphone pairs. Although this system is able to perform sound localization in three dimensions as well as resolving front-back confusions, this is only possible through the use of active audition. The restriction to a single frequency band in order to avoid phase ambiguities in the ITD computation seems to be too much of a constraint for practical applications. Additionally, extracting ITDs from several frequency bands with the method proposed in refs. [16, 23] would entail a considerable additional computational overhead. In this respect, the algorithm described here is much more robust and suitable for future extensions.

5. Conclusions and Future Work

This paper presented a bio-mimetic algorithm using ITD for sound localization with only a pair of microphones. Given the special structure of the robot, a head model was proposed first. Then, quadratic interpolation was applied to improve the spatial resolution and a novel cue selection mechanism was introduced to combat the reverberation. The algorithm was tested with the real-world data. Results showed that the algorithm is able to localize accurately a sound source in a reverberant environment with 10° resolution. The results suggested that it can simulate both localization of several concurrently active independent sources and suppression of the localization of reflected sound by considering ITD cues only when IC at the corresponding critical band is larger than a threshold. Since the number of samples required for each cross-correlation is relatively small, it is suitable for the continuous target tracking and real-time implementation.

In our current research, the overall goal is to localize a specific sound source for the robot to arrive to the object. What is shown in this paper is the ability of the system to detect and localize the sound source. The preliminary results are promising and have demonstrated the feasibility of our proposed algorithm for binaural robotic sound localization, which encourage us to continue the development and improvement of the method. The limitations of the method require further investigation. Work in progress involves extending the results to real rooms over a range of reverberation times. In the future, more experiments could be done for the cases of moving sound sources, moving robot, several simultaneous sound sources and sound localization accuracy with high-powered noise sources.

Acknowledgment

The authors are very grateful to the technical editor and anonymous reviewers for their valuable and constructive comments.

This paper was sponsored in part by the Scientific Research Foundation for the Doctoral Scholars, Liao Ning Province (Grant No: 042030) and in part supported by State Key Laboratory of Robotics and System (HIT)-SKLRS200709.

https://doi.org/10.1017/S0263574709990865 Published online by Cambridge University Press

References

1. C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two mirophones." *J. Acoust. Soc. Am.*, **108**, 1888–1905 (2000).
2. J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi and N. Sugie, "A model based sound localization system and its application to robot navigation," *Robot. Auton. Syst.* **27**, 199–209 (1999).
3. H. G. Okuno, T. Ogata and K. Komatani, "Computational Auditory Scene Analysis and Its Application to Robot Audition: Five Years Experience," *Proceedings of 2nd International Conference on Informatics Research for Development of Knowledge Society Infrastructure (ICKS'07)*, Kyoto, Japan (2007) pp. 69–76.
4. Z. Q. Wang and J. Ben-Arie, "Conveying visual information with Spatial Auditory Patterns," *IEEE Trans. Speech Audio Process.* **4**, 1092–1098 (1996).
5. J. Huang, "Spatial Sound Processing for a Hearing Robot," **In**: *Enabling Society with Information Technology* (Q. Jin, ed.) (Springer-Verlag, 2001) pp. 197–206.
6. H. Okuno, K. Nakadai and H. Kitano, "Social Interaction of Humanoid Robot Based on Audio-Visual Tracking," *Proceedings of 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2002)*, Cairns, Australia (2002) pp. 725–735.
7. D. B. Ward and R. C. Williamson, "Particle Filter Beamforming for Acoustic Sound Localization in a Reverberant Environment," *Proceedigs of the 7th International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, Orlando, FL. (2002) pp. 1777–1780.
8. J. Blauert, *Spatial hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1997).
9. S. T. Birchfiel and R. Gangishetty, "Acoustic Localization by Interaural Level Difference," *Proceedings of the 2005 International Conference on Acoustics, Speech, and Signal Processing(ICASSP05)*, Philadelphia, PA, vol. 4 (2005), pp. 18–23.
10. D. L. Wang, N. Roman and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.* **114**, 2236–2252 (2003).
11. D. Nandy and J. Ben-Arie, "Estimating the azimuth of a sound source using the binaural spectral amplitude," *IEEE Trans. Speech Audio Process.* **114**, 45–55 (1996).
12. R. A. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati and M. Williamson, "The Cog Project: Building a Humanoid Robot," **In**: *Computation for Metaphors, Analogy, and Agents* (C. Nehaniv, ed.) (Springer, 1999) pp. 52–87.
13. P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *J. Acoust. Soc. Am.* **94**, 1323–1331 (1993).
14. K. Nakadai, H. G. Okuno and H. Kitano, "Robot Recognizes Three Simultaneous Speech by Active Audition," *Proceedings of the 2003 IEEE International Conferences on Robotics and Automation (ICRA 2003)*, Taipei, Taiwan (2003) pp. 398–405.
15. K. Nakadai, D. Matsuura, H. G. Okuno and H. Tsujino, "Improvement of recognition of simultaneous speech signals using AV integration and scattering theory for humanoid robots," *Speech Commun.* **44**, 97–112, (2004).
16. E. Berglund and J. Sitte, "Sound Source Localisation Through Active Audition," *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Sytems(IROS2005)*, Edmonton, Canada (2005) pp. 653–658.
17. J. Murray, H. Erwin and S. Wermter, "Robotics Sound-Source Localization and Tracking Using Interaural Time Difference and Cross-Correlation," *Proceedings of NeuroBotics Workshop*, Ulm, Germany (2004) pp. 89–97.
18. S. G. Goodridge, Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer Interaction, ch. 3 *Ph. D. Thesis* (Springer: North Carolina State University, 1997).
19. A. A. Handzel, S. B. Andersson, M. Gebremichael and P. S. Krishnaprasad, "A Biomimetic Apparatus for Sound-source Localization," *Proceedings of 42nd IEEE Conference on Decision and Control*, Hawaii, USA (2003) pp. 5879–5884.
20. K. Nakadai, K. Hidai, H. G. Okuno and H. Kitano, "Epipolar Geometry Based Sound Localization and Extraction for Humanoid Audition," *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001)*, Maui, HI (2001) pp. 1395–1401.
21. C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.* **116**, 3075–3089 (2004).
22. E. Ben-Reuven and Y. Singer, "Discriminative Binaural Sound Localization," **In**: *Advances in Neural Information Processing Systems*, vol. 15 (S. Thrun, S. Becker and K. Obermayer, eds.) (MIT Press, Cambridge, MA, 2003) pp. 1229–1236.
23. C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. ASSP* **24**(4), 320–327 (1976).

https://doi.org/10.1017/S0263574709990865 Published online by Cambridge University Press