High performance 3D sound localization for surveillance applications

High Performance 3D Sound Localization for Surveillance Applications

Fakheredine Keyrouz and Klaus Diepold
Technische Universität München
Munich, Germany

Shady Keyrouz
Notre Dame University
Zouk Mosbeh, Libanon

Abstract

One of the key features of the human auditory system, is its nearly constant omni-directional sensitivity, e.g., the system reacts to alerting signals coming from a direction away from the sight of focused visual attention. In many surveillance situations where visual attention completely fails since the robot cameras have no direct line of sight with the sound sources, the ability to estimate the direction of the sources of danger relying on sound becomes extremely important. We present in this paper a novel method for sound localization in azimuth and elevation based on a humanoid head. The method was tested in simulations as well as in a real reverberant environment. Compared to state-of-the-art localization techniques the method is able to localize with high accuracy 3D sound sources even in the presence of reflections and high distortion.

1. Introduction

It is well known that vision is a sense that is directed, audio on the other hand is an undirected sense that helps us perceive and locate audible events outside our field of vision. The mapping of this omni-directional sensitivity from humans to humanoid robots is important, especially in cases where the robot is to survey environments in which obstructions hide the potential source of danger and might imperil the humanoid. Surveying the environment acoustically enables automatic reactions to warnings and cues of activities which are not possible based on vision alone.

Surveillance technology is becoming widely employed in many applications today ranging from domestic household appliances to industrial environments and automotive systems. It is envisaged that many audio applications for surveillance purposes such as source-position-sensing devices, cocktail-party processors, special microphones for acoustically adverse surveillance conditions will soon emerge from further utilization of surveillance technology. Towards this end, we have developed a high-resolution sound localizer inspired by the functionality of the human hearing organ as a sensitive receiver and high-resolution spectral analyzer. Using four microphones, two outside and two inside the ear canals of a humanoid head, we have built a 3D robust localization algorithm.

The ability to extract precise spatial information from the sound signals impinging at the ear drums depends on the ability to uniquely extract the direction-dependent filter shaping those signals. These direction dependent filters, known as the Head-Related Transfer Functions HRTFs, are unique and convey implicitly the direction of the sound source within their time and frequency characteristics. The head and pinnae together form a complex direction-dependent filter. The filtering action is often characterized by measuring the spectrum of the sound source and the spectrum of the sound reaching the eardrum. The ratio of these two forms the HRTF, or equivalently the head related impulse response (HRIR).

From a signal processing perspective, the underlying physical principles and a too-detailed description of a very complex system, like the ear organ of many species, are of little interest and rather undesired, because computing times are dramatically increased. Many specialized cells in the auditory pathway contribute to the highly complex signal processing, which by far exceeds the performance of modern computers. Hence, we have recently proposed a minimal-complexity sound localization system inspired by the important role of the human pinnae to focus and amplify sound [1, 2]. Based on the intelligence encapsulated within the HRTFs, which can also be interpreted as the directivity characteristics of the two pinnae [3], the model allows robots to perform localization in an indoor/outdoor environment using two synthetic pinnae and a HRTF database. The proposed algorithm deploys only two microphones and utilizes the effects of pinnae and torso on the original sound signal in order to localize one sound source in a simple matched filtering process. Extensions of the algorithm allowed 3D localization and separation of more than two concurrent sources in a real environment still using only two microphones [4].

In this paper, we present a novel monaural localization system and we combine it with the previously-proposed binaural method in order to achieve highly accurate three-dimensional localization under severe acoustical conditions.

2. Previous Work

A common approach to estimate the position of a sound source is to train a neural network to estimate the auditory

978-1-4244-1696-7/07/$25.00 ©2007 IEEE.

event from the inter-aural cues rather than to combine the cues analytically, e.g., [5]. When applying such a method, the neural network has to be trained on test material (supervised learning). The advantage of this procedure is that often very good results are achieved for stimuli that are very similar to the test material. The disadvantages are, however, the long time necessary to train the neural network and that the involved processing cannot easily be described analytically.

Recently, a biologically-based binaural technique based on a probabilistic model was proposed. The technique, [6], applies a probabilistic evaluation of a two-dimensional map containing frequency versus time-delay representation of binaural cues, a so-called activity map. However, the technique is limited to the frontal azimuthal half-plane. As for sound localization based on monaural cues, little work has been done on the subject, and few systems were able to localize sound in 3D, without becoming very complex. The localization model in [7] is based on a neuromorphic microphone that takes advantage of the biologically-based monaural spectral cues to localize sound sources in a plane. The microphone depends on a specially shaped reflecting structure that allows echo-time processing to localize the sound source.

In this paper we present a monaural localization method which extracts the HRTF from the incoming sound signal. This HRTF is then correlated with a database of HRTFs, the maximum correlation coefficient is adopted to be corresponding to the 3D sound source location. The HRTFs were measured every 5°in elevation and azimuth. An accurate, recently proposed HRTF interpolation method [8] is then used to obtain a high-spatial-resolution HRTF database with one HRTF every 1°spanning an elevation range from -20°to 60°. Each of the 28800 HRTF is 512-samples long and can be directly considered as the coefficients of a Finite Impulse Response (FIR) filter. However, for real-time processing, FIR filters of this order are computationally expensive. Applying Principal Component Analysis (PCA), the length of the HRIR was reduced to a hundred or fewer samples, considerably reducing the overall localization time and complexity. A thorough description of the PCA technique in modeling HRTFs is available in [9]. We shall denote the PCA-reduced HRTFs by H_m^{FIR}, where every HRTF has a length of m samples, and for every value of m, we have a truncated HRTF dataset.

3. Monaural System

Our proposed monaural sound localization system receives two input signals collected on two small microphones, one inserted inside and one placed outside the artificial humanoid ear.

The spatially-shaped acoustic signal inside the ear can be modeled as the original sound signal convolved with the HRTF corresponding to the target sound location. To simulate a real environment, echoes and noise are added. Hence, the signal at one of the inner microphones, the left one for instance, can be written as:

$$\mathrm{S}_{in\_L}(f) = \mathrm{S}^c_{out}(f) \cdot \mathrm{HRTF}_{ss} + \sum_{i=1}^{N} E_{in\_i}(f).\mathrm{HRTF}_i + n \tag{1}$$

where $\mathrm{S}_{in\_L}(f)$ is the signal received on the microphone inside the ear, $S^c_{out}(f)$ is the clean sound signal arriving at the ear canal, HRTF_{ss} is the correct frequency shaping response corresponding to the location of the source, $E_{in\_i}(f)$ is the i^{th} echo inside the ear arriving from some position space. The variable N represents the total number of echoes. In our case, every echo is assigned values in the interval [-20dB, -60dB]. The term HRTF_i denotes the HRTF shaping echo $E_{in\_i}(f)$. The variable n represents the noise introduced by the space and electric components.

The sound signal recorded by the microphone outside the ear, which is free of the pinnae effects, can be written as:

$$\mathrm{S}_{out\_L}(f) = \mathrm{S}^c_{out\_L}(f) + \sum_{i=1}^{N} E_{out\_i}(f) + n_s \tag{2}$$

where S_{out_L} is the signal received on the microphone outside the ear, $\mathrm{S}^c_{out\_L}(f)$ is the clean sound signal arriving at the outer microphone, $E_{out\_i}(f)$ is the i^{th} echo hitting the outside microphone. The term n_s is the noise introduced by the space.

Dividing both Equations 1 and 2, and assuming that the echo signals received are attenuated considerably, the term HRTF_{ss} dominates the division operation result. Theoretically speaking, in a noise-free anechoic environment, the division operation would result only in HRTF_{ss}.

The next step is to make a decision about the position of the sound in 3D. This is simply done by identifying the filter response that shaped the signals collected inside the ear canal. The division operation result, is sent to a bank of 28800 correlators, where it is correlated at the i^{th} correlator with the i^{th} HRTF available at the bank of correlators from an already processed lookup table. The lookup table contains the HRTFs sorted according to their azimuthal and elevation characteristics. The maximum correlation coefficient resulting from the cross-correlation between the division result and all the HRTFs is chosen to be the best estimate of the sound location. The same procedure is repeated for the right ear. The left and right blocks of Figure 1 illustrate the monaural localization at both ears.

4 Combined System

In the binaural localization case we use the system in [2]. In this context, the original signal is extracted from the received inputs, in such a way that only the HRTFs will be

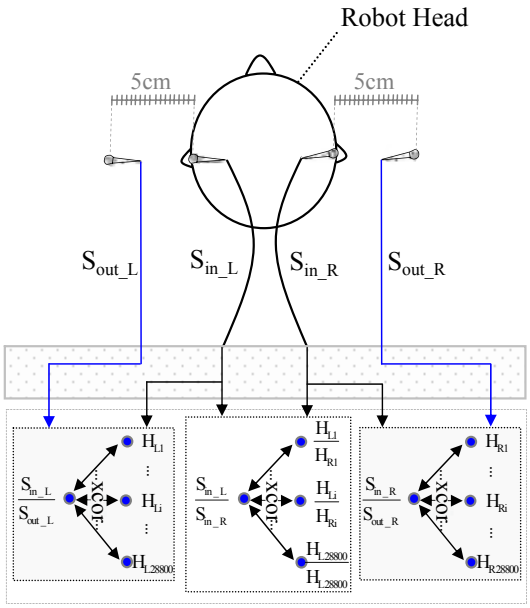

Figure 1: *Block diagram of the overall localization system.*

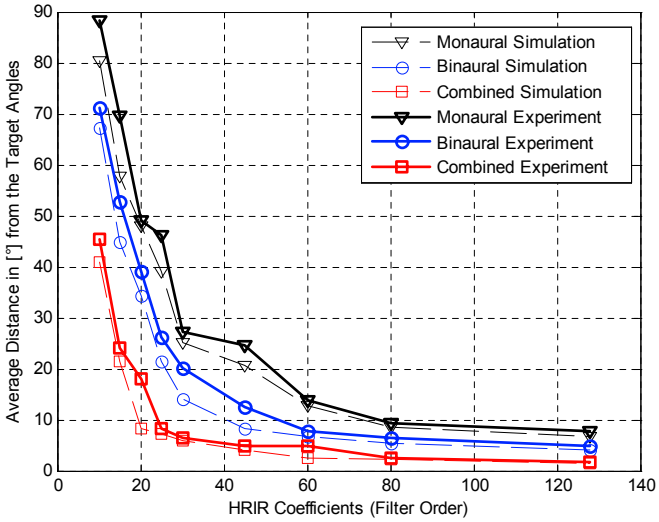

Figure 2: *Average distance, for every HRIR filter order, of the falsely localized angles with respect to their target positions*

left. The received signals at the microphones inside the ear canals are modeled as the original sound source signal convolved with appropriate HRTFs. Those signals are then divided by each other in the frequency domain. This results in the left and right HRTFs divided by each other. The sound source is canceled out and the location of the sound source is estimated by finding the maximum correlation coefficient between incoming and saved HRTF ratios. As this methods aims at canceling the effects of the incoming sound source, it is less dependent on the characteristics of the sources impinging on the artificial ears and torso, which ensures more stability and more tolerability to reverberations.The Binaural localization system is illustrated in the central block of Figure 1.

Towards achieving a better estimate of the target sound source azimuth and elevation, the 3D locations provided by both left and right monaural systems are combined with the 3D estimate given by the binaural system. In the case where two or three estimates are not more than 5°away from each other, their average is taken as the target location, and the angular error is calculated as the distance between this average and the real location. Otherwise, the angular error is calculated as the distance from the real location to the worst of the three estimates.

5. Discussion of Results

The simulation test consisted of having a 100 broadband sound signals filtered by 512-samples long HRIR at different azimuths and elevations corresponding to 100 different random source locations in the 3D space. To the simulated sound sources, white Gaussian noise and high reverberations, i.e. echoes 20dB below the signal level, were added. In order to insure rapid localization of multiple sources, small parts of the filtered left and right signals are considered (350 msecs). These left and right signal parts are then correlated with the available 28800 reduced HRIRs.

Under high reverberation conditions using the H_m^{FIR} PCA-reduced dataset, the combined system percentage of correct localization falls between 22% to 81% with the HRIR being within 10 to 45 samples, i.e. $10 \leq m \leq 45$. For a full-length HRIR of order 512, the percentage of correct localization reached 92% under the same reverberation conditions. Interestingly, for high order HRIRs, the falsely localized sound sources fall within the close neighborhood of the simulated sound source locations. A plot reporting how far, on average, are the falsely localized angles from their target location, can be seen in Fig. 2. The dashed lines and the rigid lines correspond to the simulation and experimental results, respectively. The Figure shows the performance of the monaural system (triangles), the binaural system (circles), and the combined system (squares). Intuitively, with more intelligence encapsulated within the HRIR , the localization accuracy increases. Hence, with more HRIR samples, the average distance to the target sound source location decreases. The combined system reports worst angular error of 40.83°with a HRIR order of 10, and best angular of 1.45°with a HRTF order of 128. The best performance of the binaural system was 5°compared to 6.9 °for the monaural system both operating with a HRIR order of 128.

In our household experimental setup, 100 binaural different recordings were obtained using of a broadband sound

signal, placed 2 meters away at different angle locations around the KEMAR head equipped with two small artificial ears in a highly-reverberant room. To keep a fair comparison with the simulation setup, each of the recordings was 350 msecs long, and the reverberation was kept around 20dB below the signal level. The microphones were placed inside the ears at a distance of 26 mm away from the ear's opening. The outside microphones are 5cm facing the inside ones. The recorded sound signals, also containing external and electronic noise, were used as inputs system. A HRIR database reduced using the PCA method, H_m^{FIR}, was available for the test.

The combined system percentage of correct localization falls between 6% to 74% with the HRIR being within 10 to 45-samples long, i.e. $10 \leq m \leq 45$. For a full-length HRIR, i.e. 512-samples long, the percentage of correct localization reached 81% under the same reverberation conditions. Similar to the simulations results, for high order HRIRs, the falsely localized angles fall in the vicinity of the target sound source. Figure 2 illustrates the average distance to the target angles. The combined system yielded worst angular error of 45.33°with a HRIR order of 10, and best angular of 1.6°with a HRIR order of 128. For the same order, the binaural system reported 4.2°compared to 7.8 °for the monaural system.

Furthermore, we have compared our experimental results to the method in [10]. This method uses 8 microphones and applies the TDOA algorithm to localize sound sources in three dimensions. Table 1 shows the performance of this system as compared to our system. Like in [10], the sounds we have used have a large bandwidth, e.g. fingers snapping and percussive noises. Using only 4 microphones, our system performed more accurately when localizing the sound sources situated at the same distance, azimuth and elevation angles as in [10].

Table 1: *Mean Angular Error Comparison with [10].*

Distance	**Elevation**	**Mean Error as in [10]**	**Mean Error**
3 m	-7°	1.7°	1.6°
3 m	8°	3°	1.7°
1.5 m	-13°	3.1°	1.9 °
0.9 m	24°	3.3°	2.4°

6. Conclusions

We have proposed a sound localization method which is robust to high reverberation environments and which does not require any noise cancelation schemes. The method was able to accurately localize sound sources in three dimensions using only 4 microphones. Targeting a real-time implementation on robotic platforms we have used PCA to truncate the HRTF database in such a way that fast tracking of a moving sound is achieved. The precision of the localization method is simulated and experimentally tested in a highly-reverberant environment. Compared to other localization algorithms, our system is outperforming in terms of localization accuracy and processing power.

On the other hand, the presented algorithm, cannot estimate the distance of the sound sources, and does not have the functionality of localizing or separating concurrent sound sources. Nevertheless, since the HRTFs are unique for every angle around the humanoid head, and also for every distance, using truncated HRTF databases measured at different distances from the humanoid head, is thought to enable the humanoid to perform distance estimation as well.

References

[1] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3d localization based on hrtfs," in *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006, pp. 341–344.

[2] F. Keyrouz and K. Diepold, "An enhanced binaural 3d sound localization algorithm," in *proceedings of IEEE Int. Symposium on Signal Processing and Inf. Technology (ISSPIT)*, Vancouver, Canada, 2006, pp. 663–665.

[3] J. Blauert, "An introduction to binaural technology," in *Binaural and Spatial Hearing*, R. Gilkey, T. Anderson, Eds., Lawrence Erlbaum, USA-Hilldale NJ, 1997, pp. 593–609.

[4] F. Keyrouz, W. Maier, and K. Diepold, "Robotic localization and separation of concurrent sound sources using self-splitting competitive learning," in *Proc. of the First IEEE Symp. on Comput. Intell. in Image and Signal Processing (CIISP)*, Hawaii, 2007, (to appear).

[5] F. Keyrouz, F. Lazaro-blasco, and K. Diepold, "Hierarchical fuzzy neural networks for robotic 3d sound source sensing," in *Proc. IEEE intl. Symp. on Neural Networks (ISNN)*, China, 2007, (to appear).

[6] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Krner, "A probabilistic model for binaural sound localization," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 36, no. 65, pp. 982–994, 2006.

[7] P. Chiang-Jung, J. Harris, and J. Principe, "A neuromorphic microphone for sound localization," in *Proc. IEEE/RSJ Intl. Conf. on Intell. Rob. and Sys.*, USA, 2003, pp. 1147–1152.

[8] F. Keyrouz and K. Diepold, "Efficient state-space rational interpolation of hrtfs," in *Proc. Audio Eng. Soc. (AES) 28th Intl. Conf.*, Pitea, Sweden, 2006, pp. 185–189.

[9] D. Kistler and F. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1637–1647, 1992.

[10] J. M. Valin, F. Michaud, J. Rouat, and D. Ltourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE Intl. Conf. on Intelligent Robots and Systems*, Saitama, Japan, 2003, pp. 1228–1233.