A NEW METHOD FOR BINAURAL 3-D LOCALIZATION BASED ON *HRTFs*

Fakheredine Keyrouz, Youssef Naous, Klaus Diepold

Technische Universität München
80290 Munich, Germany

ABSTRACT

A modern technique for robotic sound source detection using a dataset Head-Related Transfer Functions (HRTFs) is presented. To ensure fast detection, the HRTFs are reduced using three different techniques, namely; Diffuse-Field Equalization, Balanced Model Truncation, and Principle Component Analysis. A new criterion introduced is to be satisfied by a set of output signals from the microphones of a dummy robot head. This criterion is then used to find the sound source location in accordance with the reduced HRTF datasets. The suggested method is verified through simulated examples and further tested in a household environment. This novel technique provides estimates of azimuth and elevation angles in free space by using only two microphones. It also uses a simple algorithm compared to the more complicated algorithms used in similar localization processes.

1. INTRODUCTION

Within the context of robotic spatial sound localization systems, a rich number of models have already been proposed, most of them relying on more than two microphones to detect and track sound in a 3D space. Mathematical models of sound wave propagation and diffraction by the head and pinnae were found to significantly depend on the specific characteristics of the sources and the environment, and are therefore complex and hard to optimize. Adaptive neural network structures have also been proposed to self-adjust a sound localization model to particular environments. While these networks have been intended to work in specifically controlled milieus, they become very complex in handling multiple sources in reverberant environments. Other approaches were dedicated to mimic the human biological sound localization process by building models of the outer, middle and inner ear, and using knowledge of how acoustic events are transduced and transformed by biological auditory systems. Obviously, the difficulty with this approach is that neurophysiologists do not completely understand how living organisms localize sounds. For instance, the question of what primitive mammals like bats experience and how they process sound with only two ears and a pea-sized brain remains a major mystery.

Recently, there have been some achievements in sound localization using microphone arrays [1], [2]. Only little success, however, has been attained using two microphones. In [3], a biomimetic algorithm was proposed which determines the direction of arrival of sound by devising two curves, the acoustical phase difference and the intensity level difference between two microphones as functions of the measured frequency. These curves are then compared to a table of theoretically generated curves in order to determine the theoretical direction of the impinging sound source. However, due to the symmetrical geometry of the front and back hemispheres, the algorithm spends time distinguishing between the two hemispheres. The algorithm is limited to the bandwidth of the source and its performance deteriorates in the presence of acoustical noise and electronic noise.

In the present work, we propose an effective sound detection technique which uses two small microphones placed inside the ear canal of a robot dummy head, equipped with an artificial ear, and mounted on a torso. Matched filtering using the HRTFs is then applied. This set up proved to be very noise-tolerant and able to localize sound up to a very high precision in free space.

2. DUMMY HEAD HRTFs

The well-known Head Related Transfer Function (HRTF) [4] encapsulates both the elevation-dependent spectral filtering of sounds by the torso, head and pinnae along with two of the primary cues used for sound localization, namely, Interaural Time Difference (ITD) and Interaural Level Difference (ILD) created by sound between the two ears.

The aim is to build a robotic sound source localizer using generic HRTFs measurements, then use these measurements and develop a low-complexity model for azimuth and elevation estimation. Experiments have shown that measured HRTFs from an individual can undergo a great deal of distortion (i.e. smoothing, etc.) and still be relatively effective at generating spatialized sound. Additionally, researchers have shown that HRTFs can be relatively effective for generating spatial sound for another

person. We can take advantage of these facts in sound synthesis to greatly simplify the task of robotic sound detection by using approximations of an individual's HRTFs.

One way to approach this approximation is to use generic datasets, where we have a number of HRTF datasets available from individuals whose characteristics represented a range of those found in the general population. In this way, a listener could be fit with the HRTF dataset that most matched their own physical characteristics.

KEMAR (Knowles Electronics Manikin for Acoustic Research) is a standard based on common human anthropomorphic data and designed for making HRTF measurements [5]. The research group at MIT Media Lab has made extensive measurements using KEMAR. They provide 710 measurements collected over a wide range of spatial locations, with each HRTF having 512 samples.

The KEMAR HRTFs can be modeled as linear time-invariant digital filters. Thus, they can be presented as either Finite Impulse Response filters (FIR) or Infinite Impulse Response filters (IIR). We recap three HRTF reduction techniques, two FIR and one IIR, that are implemented on the KEMAR dataset, leading to less complicated datasets of HRTFs. Using these datasets, we present a novel approach to localize sound using only two microphones in real 3-D space, as opposed to conventional methods utilizing HRTFs to synthesize sound for virtual reality.

3. MODEL REDUCTION TECHNIQUES

The KEMAR HRTF dataset contains the impulse response of the actual filter, thus, the 512 samples can be directly considered to be the FIR representation of the filter. However, such real-time FIR filters of this order are computationally expensive. Moreover, this dataset is to be used to perform localization of sound sources and to account for head movements, which implies that the dataset has to be stored to insure rapid switching between HRTFs. Therefore, using the 512 samples slows down the localization process and does not offer memory saving.

3.1. Diffuse-field equalization

In this technique, we reduce the HRTF length from 512 samples to 128 samples. Thus, a 128-order FIR filter is used to model the KEMAR HRTF dataset [6]. The following steps are used to reduce the sample length:

- initial time delay is removed from the beginning of the response, which is around 10-15 samples
- features that are independent of the incident angle, currently being modeled, are removed [6]

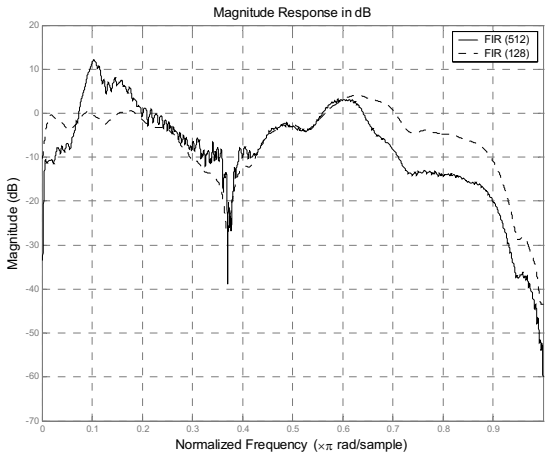

Fig. 1. Magnitude response of the original 512-FIR (solid) and the reduced 128-FIR (dashed) of an HRTF (left ear, 0º azimuth)

- smoothing of the magnitude response using a critical-band auditory smoothing technique [7]
- a minimum-phase filter is constructed

The final outcome is a 128-sample response, which can be considered as an FIR filter representation of the HRTF. Figure 1 shows the diffuse-field equalized HRTF filter response in comparison to the original 512-sample HRTF.

3.2. Balanced model truncation

BMT is used to design a low-order IIR filter model of the HRTF from a high-order FIR filter response. A detailed description of the BMT technique is available in [8], however; a brief outline will be presented here. BMT analyzes a linear time-invariant system in order to locate which states are contributing to its input/output response. Once these states are located, the system is truncated and only the selected states are used to model the filter itself.

This technique builds on the previous one, that is, we consider the 128-samlpe FIR filter. This filter can be written as: $F(z) = c_0 + c_1 \cdot z^{-1} + c_2 \cdot z^{-2} + ... + c_n \cdot z^{-n}$, where n = 127. This filter can be represented as state-space difference equations:

$$x(k+1) = A \cdot x(k) + B \cdot u(k)$$
$$y(k) = C \cdot x(k) + D \cdot u(k)$$

Then, a transformation matrix T is found such that the controllability and observability grammians P and Q are equal and diagonal, thus the system is balanced, and the system states are ordered according to their contribution to the system response. The order of the states is reflected in the Hankel Singular Values (HSV) of the system. Thus, the balanced system can be divided into two sub-systems: the truncated system of order *m*, where the first *m* HSVs are used to model the filter, and the rejected system of order (*n-m)*. Figure 2 shows the BMT-reduced IIR (25^{th} order) in comparison to the FIR (128^{th} order).

Fig. 2. Magnitude response of the 128-FIR (solid) and the reduced 25-IIR (dashed) of an HRTF (left ear, 0º azimuth)

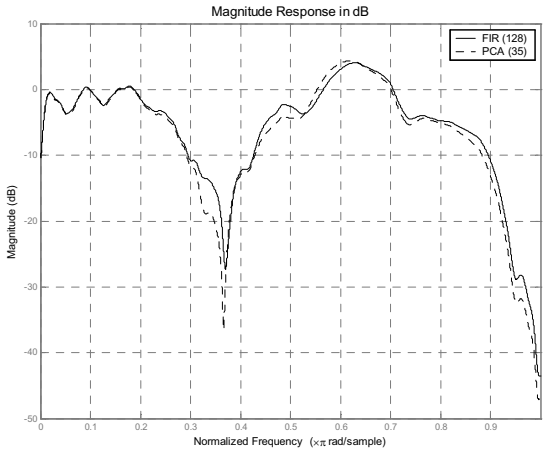

Fig. 3. Magnitude response of the 128-FIR (solid) and the reduced 35-FIR (dashed) of an HRTF (left ear, 5º azimuth)

3.3. Pricipal component analysis

PCA is used to reduce the number of samples required to represent each 128-sample diffuse-field equalized HRTF. A thorough description of the PCA technique in modeling HRTFs is available in [9]. The PCA aims at minimizing the amount of storage space needed for the HRTF dataset by selecting *m* representatives from the whole dataset. Figure 3 shows the PCA-reduced FIR (35^{th} order) in comparison to the FIR (128^{th} order).

4. PROPOSED LOCALIZATION TECHNIQUE

Our proposed approach for sound localization depends on finding the inverse filters of the whole reduced HRTF dataset, and then saving these inverses to use them in localization. This process is to be done offline, that is, we calculate the inverse filter response of the HRTF dataset and save it for future use. The inverse filter is directly available by simply exchanging the values of the numerator and the denominator, that is, for the FIR filter: $F(z) = c_0 + c_1 \cdot z^{-1} + c_2 \cdot z^{-2} + ... + c_n \cdot z^{-n}$, the inverse filter will be: $G(z) = \dfrac{1}{c_0 + c_1 \cdot z^{-1} + c_2 \cdot z^{-2} + ... + c_n \cdot z^{-n}}$.

Similarly, the same concept applies for the IIR filter.

Our localization scenario is the following: multiple sound sources are generating signals; these signals are filtered with a corresponding HRTF, depending on the location of the sound source. In order to insure rapid localization of multiple sources, a small part of the filtered signal is considered (about 350msec). This signal part is then used to convolve with the available reduced HRTF set. Since the original signals arrive in pairs, that is, one representing the left ear and the other representing the right ear, we can simply calculate the correlation factor between the two resulting outputs. Theoretically, this correlation factor should yield a value close to one, since the two signals are supposed to be almost the same. Therefore, we base our localization on the maximum correlation factor obtained. Moreover, to insure an accurate localization decision, the minimum distance measure is also calculated. Theoretically, the distance between the two signals (left and right) should yield a minimum value since the two signals are supposed to be almost equal. The following is a flow of the algorithm:

$$
\begin{aligned}
&y_L(t) = \textit{received signal at left ear} \\
&y_R(t) = \textit{received signal at right ear} \\
&x_L{}^{(i)}(t) = y_L(t) * H^{-1}{}_L{}^{(i)} \quad i = 1 \to n \\
&x_R{}^{(i)}(t) = y_R(t) * H^{-1}{}_R{}^{(i)} \quad i = 1 \to n \\
&n = \textit{number of HRTFs in the dataset} \\
&c^{(i)} = corr(x_L{}^{(i)}(t),\ x_R{}^{(i)}(t)) \\
&d^{(i)} = \sum_m (x_L{}^{(i)}(t) - x_R{}^{(i)}(t))^2 \\
&m = \textit{rank of reduced HRTF}
\end{aligned}
$$

The above localization technique is applied using the previously described reduced HRTF models.

5. SIMULATION AND EXPERIMENTAL RESULTS

The simulation test consisted of having a sound signal filtered out by the effect of the 512-sample HRTF at a certain azimuth and elevation. Thus, the test signal was virtually synthesized using the original HRTF set. The reduction techniques, namely; diffuse-field equalization, BMT, and PCA were used to create three different reduced models of the original HRTFs.

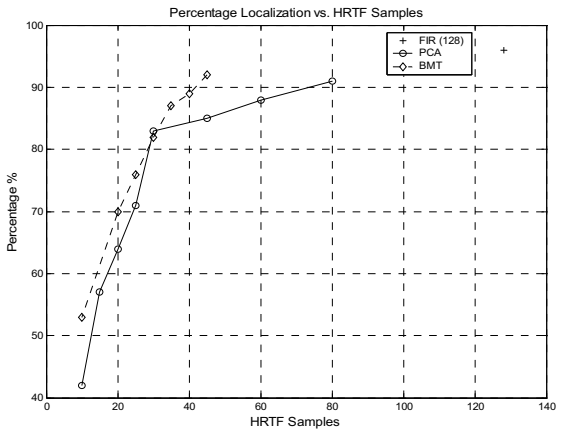

Fig. 4. Percentage of correct localization using reduced HRTFs

Using the diffuse-field equalized HRTF set, the simulated percentage of correct localization was around 96%, whereas; using the BMT-reduced set, the localization percentage was between 53% to 92% with the HRTF being within 10 to 45 samples, moreover, the PCA-reduced set yielded a correct localization of 42% to 91% with the HRTF dataset being represented by 10 to 80 bins. Figure 4 shows the simulated outcome of the proposed localization technique. Most significantly, it was observed that, up to 30 bins PCA-reduced and 35 BMT-reduced HRTFs, all the falsely localized angles fall within the very close vicinity of the original sound locations.

On the other hand, several binaural recordings from different directions were obtained using a dummy head and torso with two artificial ears in a reverberant room. The microphones were placed near the ear drums at a distance of 26 mm away from the ear's opening. The recorded sound signals, also containing external and electronic noise, were used as inputs to our localization algorithm. Only 30 bins of PCA-reduced and 35 bins of BMT-reduced HRTFs were used. All the estimated azimuth and elevation angles turned out to be either at or in the very near neighborhood of the original angles. This close localization is due to the difference between the dummy model used in the experiment and the KEMAR model used to obtain the HRTF dataset.

6. CONCLUSIONS

In this paper, we have presented a simple yet promising sound source localization method using an HRTF database. From the above result, it is clear that the proposed localization method has the ability of precise azimuth and elevation estimation. However, since the HRTF dataset is obtained on the horizontal plane from 0° to 180° with 5° increment and on the vertical plane from -40° to 90° with 10° increment, the above result indicates that we can find the source location with an accuracy of several degrees. That is to say, we can localize the sound source with an accuracy of about 5° and if we construct the HRTF dataset with a less increment, the resolution of estimation will be increased.

Moreover, the proposed method was demonstrated to localize free space sound using two microphones. Conventional methods such as the ITD method cannot find the azimuth and elevation of a sound source by using only two microphones, without becoming very complex. The proposed technique utilized a very simple algorithm. This simplicity, compared to the complexity of current localization algorithms, provides an easy implementation for robot platforms and allows for a fast localization of multiple sources. For future work, the proposed detection algorithm will be expanded to perform sound range estimations.

7. REFERENCES

[1] O. Jahromi and P. Aarabi, "Theory and Design of Multirate Sensor Arrays", IEEE Transactions on Signal Processing, vol. 53, no. 5, pp. 1739–1753, May 2005.

[2] Q.H. Wang, T. Ivanov, and P. Aarabi, "Acoustic Robot Navigation Using Distributed Microphone Arrays", Information Fusion (Special Issue on Robust Speech Processing), vol. 5, no. 2, pp. 131–140, June 2004.

[3] A.A. Handzel and P.S. Krishnaprasad, "Biomimetic Sound-Source Localization", IEEE Sensors Journal, vol. 2, no. 6, pp. 607 – 616, Dec. 2002.

[4] J. Blauert, "Spatial Hearing: The Psychophysics of Human Sound Localization", rev. ed. MIT-Press, Cambridge, MA, 1997.

[5] W. G. Gardner and K. D. Martin, "HRTF Measurements of a KEMAR", J. Acoust. Soc. Amer., vol. 97, no. 6, pp. 3907–3908, 1995.

[6] H. Møller, "Fundamentals of Binaural Technology", Appl. Acoust., vol. 36, no. 3-4, pp. 171–218, 1992.

[7] J. Mackenzie, J. Huopaniemi, et al. "Low-Order Modeling of Head-Related Transfer Functions Using Balanced Model Truncation", IEEE Signal Processing Letters, vol. 4, no. 2, pp. 39-41, Feb. 1997.

[8] B. Beliczynski, I. Kale, and G.D. Cain, "Approximation of FIR by IIR Digital Filters: An Algorithm Based on Balanced Model Reduction", IEEE Trans. Signal Processing, vol. 40, no. 3, pp. 532–542, Mar. 1992.

[9] D.J. Kistler and F.L. 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, Mar. 1992.