An Enhanced 3D Sound Localization Algorithm Using HRTFs

Fakheredine Keyrouz and Klaus Diepold
Technische Universität München
80290 Munich, Germany

***Abstract*— For sound localization methods to be useful in real-time scenarios, the processing power requirements must be low enough to allow real time processing of audio inputs. we propose a new binaural sound source localization technique based on using only two microphones placed inside the ear canal of a robot dummy head. The head is equipped with artificial ears and is mounted on a torso. In contrast to existing 3D sound source localization methods using microphone arrays, the novel method presented employs only two microphone and is based on a simple correlation approach using a generic set of HRTFs. The proposed method is demonstrated through simulation and is further tested in a household environment. This set up proves to be very noise-tolerant and is able to localize sound sources in free space with high precision.**

I. INTRODUCTION

In the field of acoustic, determining the direction from which a sound is coming in 3D space has been approached in several ways [1], [2], [3]. Using standard techniques, a large array of microphones can be used to localize sound sources in a three-dimensional space.

However, most humans and other mammals use binaural hearing to be able to accurately localize sound with only two ears. From the viewpoint of signal processing, the human hearing organ is a signal processor par excellence. Binaural hearing not only has the capabilities to concentrate on one sound source in a crowd of concurrent sound sources and discriminate between the different sources, but also is able to suppress noise, reverberance and sound colouration to a certain extent [4]. When placed in a free sound field, a listener will obstruct an incoming sound wave. The external sound field has to couple into the listener's ear canals. The relative positions of the two ear canals and the sound sources lead to a coupling that is strongly dependent on frequency, expect at very low frequencies. In this context not only the two pinnae but also the whole head and the torso have an important functional role, which is best described as a spatial filtering process. This linear filtering is usually quantified in terms of the HRTFs. In the general definition of the HRTF all linear properties of the sound transmission are included. All proposed descriptors of localization cues, such as inter-aural differences in arrival-time, ITD, in phase, IPD, in level/intensity, ILD/IID as well as monaural cues, are contained in the HRTFs. They can thus be derived from the HRTFs, whereas the opposite is not generally the case [4].

More challenging for sound source detection systems, is the ability of humans to localize sounds monaurally. A permanent process of interpretation seems to be an essential part of hearing, since hearing using only one ear cannot be explained on the basis of binaural cues like ITDs and ILDs in the input stimuli. This signifies that the interpretation of spectral cues are very essential for localization and that the extraction of such cues must be explained within the framework of any localization model.

In a similar fashion to the decoding process which the auditory system performs when transforming the two one-dimensional signals at the eardrums back into a three-dimensional space representation, it has been suggested that robotics can benefit from the intelligence encapsulated within the HRTFs to localize sound in 3D [5]. Motivitated by the important role of the human pinnae to focus and amplify sound, and since the HRTFs can also be interpreted as the directivity characteristics of the two pinnae [4], a robot should perform sound localization in three dimensional space using only two microphones, two synthetic pinnae and a HRTF database. The previously proposed system in [5] utilizes the effects of pinnae and torso on the original sound signal in order to localize the sound source in a simple matched filtering process. In this paper, a new localization algorithm is presented which produces better results and greatly reduces the processing requirements compared to the previously proposed algorithm.

II. HRTFS REDUCTION TECHNIQUES

Our goal is to build a binaural sound source localizer using a set of generic HRTF measurements. We use these measurements and develop a low-complexity model based on simple correlation for estimating the azimuth and the elevation for an impinging sound wave. Experiments have shown that measured HRTFs from an individual can undergo a great deal of distortion (i.e. smoothing, reduction, etc.) and still be relatively effective at generating spatialized sound (Blauert, 1997). This implies that the reduced HRTF still contain all the necessary descriptors of localization cues and is able to uniquely represent the transfer of sound from a particular point in the 3D space. We can take advantage of this fact to greatly simplify the task of sound source localization by using approximations of an individual's HRTFs, shortening thus the length of each HRTF and consequently reducing the overall localization processing time.

One public set of HRTFs are those collected at MIT. Using the KEMAR (Knowles Electronics Manikin for Acoustic

Research) which is a standard manikin based on common human anthropomorphic data, the research group gathered 710 accurate measurements taken over a broad range of spatial locations, with each HRTF having a length of 512 samples.

The KEMAR HRTFs can be modeled as a set of linear time-invariant digital filters, being represented either as Finite Impulse Response (FIR) filters or as Infinite Impulse Response (IIR) filters. We investigate three techniques for reducing the length of the HRTF, two FIR and one IIR, which are applied to the KEMAR dataset, and which lead to a significant reduction in the size of the measured HRTF dataset. Using the reduced dataset, we present a novel approach to localize sound sources using only two microphones in a real environment.

The KEMAR dataset contains the impulse responses of the actually measured HRTF filters. The 512 samples of each HRTF-measurement can directly be considered to be the coefficients of a FIR representation of the filter. However, for real-time processing FIR filters of this order are computationally expensive. Moreover, the 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 allow for fast switching between HRTFs. Using the 512 samples slows down the localization process and does not offer memory savings. The original KEMAR HRTFs containing the 512 coefficients of the FIR filter will be denoted by H_{512}^{FIR}.

A. Diffuse-Field Equalization

Our goal is to shorten the length of the original filters in order to reduce the computational burden for convolution, while preserving the main characteristics of the measured impulse responses. To this end, we adopt the algorithm proposed by [6] for a diffuse-field equalization (DFE). In DFE, a reference spectrum is derived by power-averaging all HRTFs from each ear and taking the square root of this average spectrum. Diffuse-field equalized HRTFs are obtained by de-convolving the original by the diffuse-field reference HRTF of that ear. This leads to the fact that the factors that are not incident-angle dependent, such as the ear canal resonance, are removed. The DFE is achieved according to the following four steps: 1) Remove the initial time delay from the beginning of the measured impulse responses, which typically has a duration of about 10-15 samples. 2) Remove features from modeling that are independent of the incident angle [6]. 3) Smooth the magnitude response using a critical-band auditory smoothing technique [7]. 4) Construct a minimum-phase filter, ensuring thus stability for the final filter and its inverse.

This way we shorten the length of the FIR representation of the original KEMAR HRTFs, H_{512}^{FIR}, from 512 to 128 coefficients. The resulting HRTF database is denoted as H_{128}^{FIR}.

B. Balanced Model Truncation

In order to examine to which extent the HRTF can be reduced while still preserving the characteristic information which makes it unique, we reduce the previously derived diffused-field HRTF dataset further by adopting the balanced model truncation (BMT) technique 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 given in [8]. In a brief description, we determine a linear time-invariant state-space system, which realizes the filter H_{128}^{FIR}. We start representing the 128-coefficient FIR filter H_{128}^{FIR} as state-space difference equations, then a transformation matrix is found such that the controllability and observability Grammians are equal and diagonal. This is the characteristic feature of a balanced system. The corresponding 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 $m < n$, where the first m HSVs are used to model the filter, and the rejected system of order $(n - m)$. For every value of m we have a truncated HRTF dataset denoted as H_{m}^{IIR}.

C. Principal Component Analysis

As an alternative to the previous BMT approach a Principal 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, the obtained HRTF database is denoted by H_{m}^{FIR}.

III. Previous Sound Source Localization Technique

We now recollect in detail the localization method which was suggested in [5]. The main idea in this algorithm was to first minimize the HRTFs and remove all redundancy. The resulting minimized HRTFs are then used for localizing sound sources in the same way the full HRTFs would be used. The algorithm relies on a straight-forward matched filtering concept.

We assume that we have received the left and right signals of a sound source from a certain direction. The received signal to each ear is therefore the original signal filtered through the HRTF that corresponds to the given ear and direction.

Match Filtering the received signals through the correct HRTF should give back the original mono signal of the sound source. Although the system has no information about what the sound source is, the result of filtering the left received signal by the correct inverse left HRTF should be identical to the the right received signal filtered by the correct inverse right HRTF.

In order to determine the direction from which the sound is arriving, the two signals must be filtered by the inverse of all of the HRTFs. The inverse HRTFs that result in a pair of signals that closely resemble each other should correspond to the direction of the sound source. This is determined using a simple correlation function. The direction of the sound source is assumed to be the HRTF pair with the highest correlation. This method is illustrated in Figure 1.

Due to the computational complexity of filtering in time domain, the algorithm is applied in the frequency domain. The signals and reduced HRTFs are all converted using the Fast Fourier Transform (FFT). This changes all the filtering operations to simple array multiplications and divisions.

We assume that the reduced HRTFs, i.e. using diffused, BMT and PCA, have already been calculated and saved. Once the audio samples are received to the left and right inputs, they must also be transformed using FFT. Then, the transformed signal is divided (or multiplied by a pre-calculated inverse) by each of the HRTFs. Finally, the correlation of each pair from the left and right is calculated. There are 1420 array multiplications, 1420 inverse Fourier transforms, and 710 correlation operations. After the correlations are found, the direction that corresponds to the maximum correlation value is taken to be the direction from which the sound is arriving.

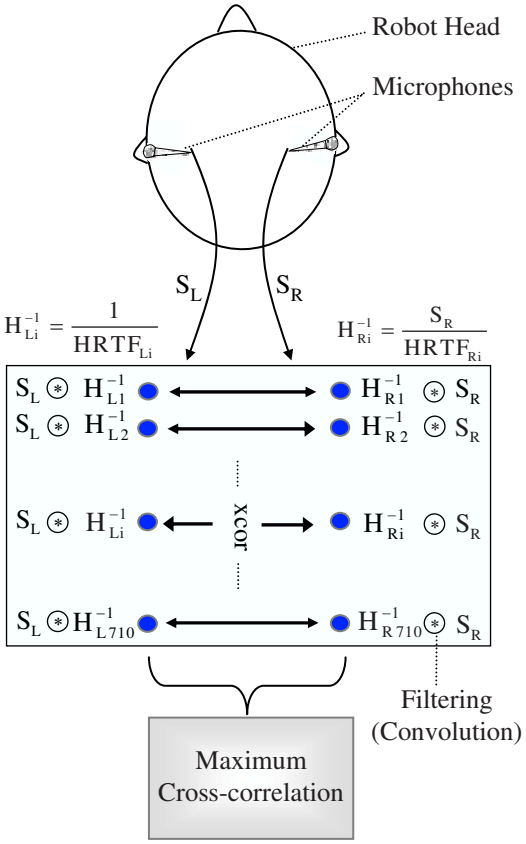

Fig. 1. Flowchart of the sound localization algorithm as proposed in [5]

IV. Source Cancelation Algorithm

In the previous algorithm, the main goal was to pass the received signal through all possible inverse filters. The set of filters from the correct direction would result in canceling the effects of the HRTF and extracting the original signal from both sides.

However, a more direct approach can be taken to localize a sound source. Instead of attempting the retrieve it, discarding the original signal from the received inputs so that only the HRTFs are left may be possible. Such an approach is denoted as the Source Cancelation Algorithm (SCA) and is illustrated in Figure 2. Basically, the received signals at the microphones

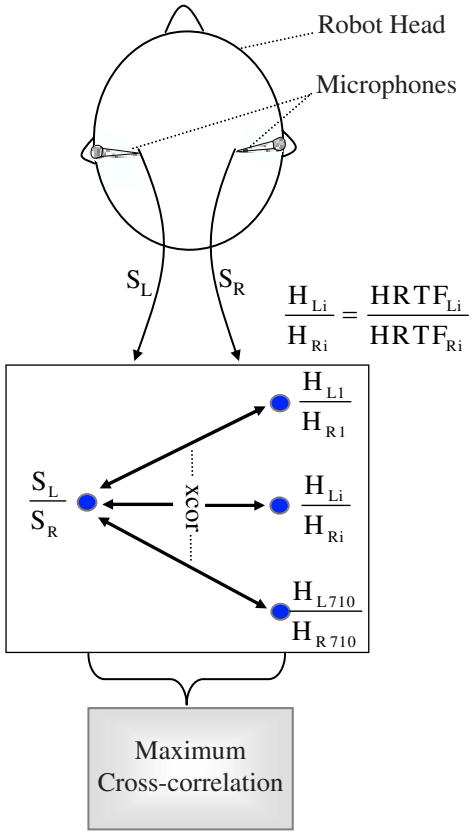

Fig. 2. Flow chart of the SCA sound localization algorithm.

inside the ear canals could be reasonably modeled as the original sound source signal convolved by the HRTF. Looking at the signals in frequency domain, we see that if we divide the left and right transformed signals, we are left with the left and right HRTFs divided by each other. The sound source is canceled out.

With two Fourier transforms and one array division operation, the original signal is removed and the HRTFs are isolated. The resulting ratio can then be compared to the ratios of HRTFs which are stored in the system. These ratios are assumed to be pre-calculated offline and saved in the system database, since they do not change. Additionally, the correlation operation is performed in the frequency domain to eliminate the need for inverse Fourier transforms.

In a hardware-based application, using the Source Cancelation Algorithm would greatly reduce hardware complexity as well as speed up processing. Compared to the original algorithm, this new approach eliminates 1420 array multiplications and 1420 inverse Fourier transforms, and replaces them with one single array multiplication.

V. Simulation and Experimental Results

The simulation test consisted of having a broadband 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. For the test signal synthesis, a total of 100 random HRTFs were used corresponding to 100 different random source location in the 3D space. In order to insure rapid localization of multiple sources, small parts of the filtered left and right signal is considered (about 350msec). These left and right signal parts are then transformed using FFT and divided, the division result is then correlated with the available 710 reduced HRTF ratios, i.e. $\frac{\text{HRTF}_{Li}}{\text{HRTF}_{Ri}}$. Basically, the correlation should yield a maximum value when the saved HRTF ratio corresponds to the location from which the simulated sound source is originating. Therefore, we base our localization on the obtained maximum for the correlation factor. The reduction techniques, namely, diffuse-field equalization, BMT, and PCA were used to create three different reduced models of the original HRTFs. The performance of each of these models under the SCA is illustrated in Figure 3. The solid lines and the star sign in the Figure shows the SCA percentage of correct localization versus the length of the HRTF in samples. For comparison, the dashed lines and the plus sign in the figure refer to the previous method performance [5]. Using

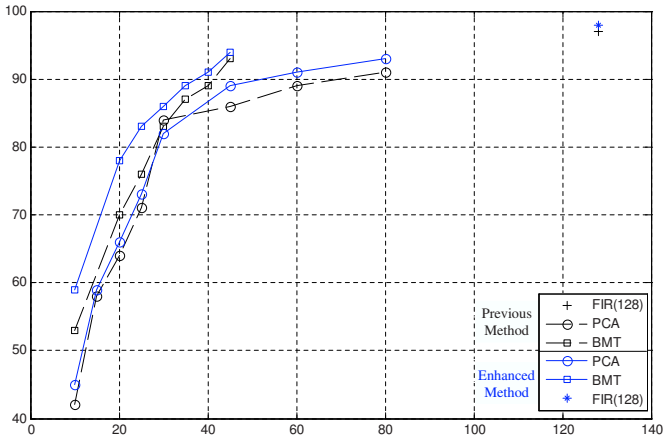

Fig. 3. Percentage of correct localization using SCA compared with the method in [5]

the diffuse-field equalized 128 samples HRTF set, the SCA simulated percentage of correct localization is around 97%, this means that out of 100 locations, 97 where detected by our SCA algorithm compared to 96% for the previous method. Using the BMT-reduced set, the SCA localization percentage was between 59% to 93% compared to 53% to 92% for the previous method, with the HRTF being within 10 to 45 samples. The PCA-reduced set yielded a correct localization of 45% to 92% compared to 42% to 91% for the previous method, with the HRTF dataset being represented by 10 to 80 samples. It should be noted that, while using 30 samples PCA-reduced and 35 BMT-reduced HRTFs, all the falsely localized angles fall within the close neighborhood of the simulated sound source locations.

In our household experimental setup, 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 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 SCA localization algorithm. Only 30 samples of PCA-reduced and 35 samples of BMT-reduced HRTFs were used. All the estimated azimuth and elevation angles turned out to be either at or in the vicinity of the target angles. Due to the differences between the dummy manikin model used in the experiment and the KEMAR model used to obtain the HRTF dataset, some angles are not exactly detected at the target location where they originated from.

VI. Conclusion

We have addressed the binaural sound source localization problem. We proposed an efficient sound source localization method that demonstrates the ability of precise azimuth and elevation estimation, using a generic HRTF database. The HRTF dataset is measured on the horizontal planes from 0°to 360°with a minimum of 5°increments and on the vertical plane from -40°to 90°with 10°increments. Therefore, the results indicate that we can localize the sound source with an accuracy of about 5°. If we construct the HRTF dataset with smaller increments, the resolution of estimation will be increased. Additionally, the new algorithm was developed to further decrease the computational requirements of the method in [5]. Compared to the previous method, the SCA algorithm is able to achieve remarkable reduction in the processing requirements while increasing the accuracy of the sound localization. The efficiency of the new algorithm, suggests a cost-effective implementation for robot platforms and allows for a fast localization of multiple sources. Using the presented method, many venues for future work are to be considered, mainly range estimation and robotic monaural localization.

References

[1] A. A. Handzel, "High acuity sound-source localization by means of a triangular spherical array," in *proceedings of IEEE ICASSP*, 2005, pp. 1057–1060).

[2] H. Nakashim and T. Mukai, "3d sound source localization system based on learning of binaural hearing," in *IEEE Int. Conf. on Systems, Man and Cybernetics (IEEE SMC 2005)*, 2005, pp. 3534–3539.

[3] Y. Rui and D. Florencio, "New direct approaches to robust sound source localization," in *proceedings of IEEE ICME*, 2003, pp. 737–740.

[4] 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.

[5] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3d localization based on hrtfs," in *proceedings of IEEE ICASSP*, May 2006, (to appear).

[6] H. Mller, "Fundamentals of binaural technology," *Appl. Acoust.*, vol. 36, no. 3-4, pp. 171–218, 1992.

[7] J. Mackenzie, J. Huopaniemi, V. Vlimki, and I. Kale, "Low-order modeling of head-related transfer functions using balanced model truncation," *IEEE Singal Processing Letters*, vol. 4, no. 2, pp. 39–41, 1997.

[8] B. Beliczynski, I. Kale, and G. Cain, "Low-order modeling of head-related transfer functions using balanced model truncation," *IEEE Trans. Signal Processing*, vol. 40, no. 3, pp. 532–542, 1997.

[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.