HRTF Measurements With Recorded Reference Signal

(PREPRINT)

Marko Đurković[1], Florian Sagstetter, and Klaus Diepold[1]

[1] *Institute for Data Processing, Technische Universität München, 80290 München, Germany*

Correspondence should be addressed to Marko Đurković (durkovic@tum.de)

ABSTRACT

Head-Related Transfer Functions (HRTFs) are used for adding spatial information in 3D audio synthesis or for binaural robotic sound localization. Both tasks work best when using a custom HRTF database that fits the physiology of each person or robot. Usually, measuring HRTFs is a time consuming and complex procedure that is performed with expensive equipment in an anechoic chamber. In this paper we present a method that enables HRTF measurement in everyday environments by passively recording the surroundings without the need to actively emit special excitation signals. Experiments show that our method captures the HRTF's spatial cues and enables accurate sound localization.

1. INTRODUCTION

Head-Related Transfer Functions (HRTFs) [2,3] describe how sound is spectrally affected by the head and ears before it enters the ear canal. HRTFs differ from person to person, since they are influenced by physical properties like the size of the head or the form of the pinna.

Additionally, the spectral distortions are dependent on the direction of the sound source in respect to the listener. They can be used for adding spatial information in 3D audio synthesis or for binaural robotic sound localization. HRTFs are obtained in a time consuming and complex measurement procedure, that is generally performed with expensive equipment in an anechoic chamber.

Several different research projects, like the CoTeSys research cluster (www.cotesys.org), are working on cognitive robots that perceive their environment through a number of sensors and thus different modalities. Due to its omni-directional properties the sense of hearing is an important complement to field of view limited senses like vision. In a binaural setup the robot is equipped with two ears and microphones in each ear canal. With binaural sound localization techniques [6, 7, 10] the robot is able to determine the source direction of a sound. The accuracy of those methods is dependent on the quality of the robot's HRTF database. Even small modifications of a robot's hardware can change its HRTFs and have a negative impact on the robots localization accuracy.

HRTFs for persons or humanoid robots are measured by placing microphones in the ear canals of the respective subject. The Head-Related Impulse Response (HRIR), which is the time-domain equivalent of a HRTF, can be measured in an anechoic chamber by exciting an impulse. Due to the low energy of an impulse and the resulting poor SNR, better excitation signals have been investigated in literature. Gardner and Martin [5] used Maximum Length Sequences to measure a KEMAR's HRTFs at MIT, Algazi et al. [1] used Golay Codes to create the CIPIC database of human subjects. Farina [4] proposed exponential sweeps as excitation signals. To obtain a HRTF database these excitation signals are presented to the subject from every direction that should be present in the database and the HRIRs are calculated by deconvolution.

Research robots are often under heavy development and their technical setup and thus the robot's HRTF databases are constantly subject to change. Therefore HRTF measurements would have to be performed for every individual robot and have to be repeated each time the setup of the robot would change. To be able to quickly adapt to changes in the robot hardware we have designed a measuring process that can be done in any environment without the need for an anechoic chamber and that works without being dependent on special excitation signals. It is a modification of the traditional measurement approach and instead of measuring the HRIRs exactly, we try to ensure that the HRTF-inherent spatial cues are captured precisely. We call our technique HRTF Measurement with Recorded Reference Signal (HMRR).

2. THEORETICAL BACKGROUND

Before we present our algorithm, we will introduce the necessary theoretical concepts in this section.

2.1. Conventional HRTF measurement approach

We created several reference HRTF databases with exponential sweeps [4] as excitation signals to benchmark the results of our HMRR approach. In laboratories HRTFs are measured as exactly as possible over the whole frequency spectrum. Exponential sweeps are sinusoidal signals starting at a frequency ω_1 and exponentially rising to ω_2 during the time T. With the so called slew-rate

$$c = \frac{ln(\frac{\omega_2}{\omega_1})}{T} \tag{1}$$

an exponential sweep can be calculated by

$$s(t) = sin\left(\frac{\omega_1}{c} \cdot \left(e^{ct} - 1\right)\right), t \in [0, T]. \tag{2}$$

The excitation signal is presented by a loudspeaker in an anechoic chamber from different positions and the ear signals $x_1(t)$ and $x_2(t)$ for the left and right ear are measured. These signals can be described as the source signal convolved with an unknown impulse response $h_1(t)$ and $h_2(t)$:

$$x_i(t) = \int_{-\infty}^{\infty} s(\tau)h_i(t-\tau)d\tau = s(t) * h_i(t),\ i \in \{1, 2\} \tag{3}$$

The inverse of an exponential sweep in regards to convolution is given by

$$x_{inv}(t) = x(-t)e^{-ct} \tag{4}$$

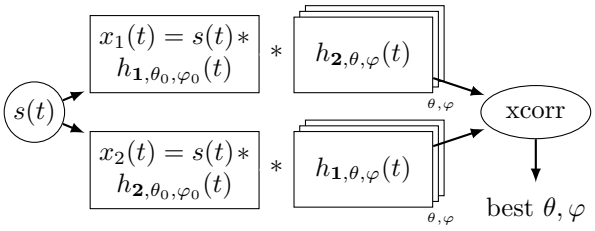

Fig. 1: Schematic view of the cross-convolution localization approach.

and finally the HRIRs $h_i(t)$ and HRTFs $H_i(t)$ for one direction can be obtained by:

$$\begin{aligned} h_i(t) = &\quad s_{inv}(t) * x_i(t), &\quad i \in \{1, 2\} \\ H_i(f) = &\quad \mathcal{F}(h_i(t)), &\quad i \in \{1, 2\} \end{aligned} \tag{5}$$

By using this method in a room with reverberations we also capture information about the room impulse response with each HRTF.

2.2. Cross-Convolution Localization

In Section 5 we measure the quality of a HRTF database with the localization accuracy. For localization we use the cross-convolution localization algorithm [10]. A schematic view of the algorithm is given in Figure 1. This algorithm takes the respective HRTF database and the recorded ear signals and returns the azimuth angle φ and elevation angle θ of the sound source.

The algorithm takes the ear recordings $x_i(t),\ i \in \{1, 2\}$ of an unknown source $s(t)$ and convolves them with each HRTF pair from the database. Each recording is convolved with the HRTF from the opposite ear, hence the term cross-convolution. If the correct HRTF pair is chosen from the database, the resulting signals $\tilde{s}_{i,\theta,\varphi}(t),\ i \in \{1, 2\}$ will be identical in an ideal recording scenario:

$$\begin{aligned} \tilde{s}_{1,\theta,\varphi}(t) &= h_{2,\theta,\varphi}(t) * x_1(t) \\ &= h_{2,\theta,\varphi}(t) * h_{1,\theta_0,\varphi_0}(t) * s(t) \\ &= h_{1,\theta,\varphi}(t) * h_{2,\theta_0,\varphi_0}(t) * s(t) \\ &= h_{1,\theta,\varphi}(t) * x_2(t) \\ &= \tilde{s}_{2,\theta,\varphi}(t) \iff \theta = \theta_0, \varphi = \varphi_0. \end{aligned} \tag{6}$$

Due to background noises and recording equipment Equation (6) will not hold true in a real case. Therefore, we choose the HRTF that maximizes the cross-correlation between the two signals $\tilde{s}_{1,\theta,\varphi}(t)$ and

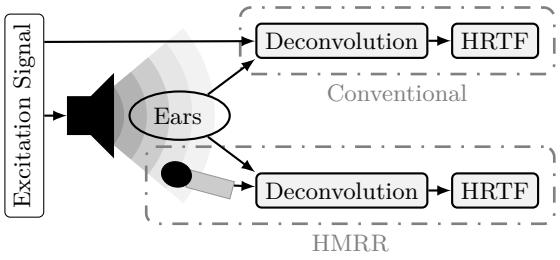

Fig. 2: Schematic view of the HRTF recording approaches. In contrast to the conventional approach HMRR does not need information about the excitation signal in addition to the ear recordings.

$\tilde{s}_{2,\theta,\varphi}(t)$:

$$(\hat{\theta},\hat{\varphi}) = \arg\max_{\theta,\varphi} \hat{s}_{1,\theta,\varphi} \star \hat{s}_{2,\theta,\varphi} \tag{7}$$

The biggest advantage of the cross-convolution approach is that no deconvolution or spectral inversion is required.

3. HMRR APPROACH AND ALGORITHM

We want to be able to record HRTFs in arbitrary rooms without generating and emitting special test signals. We are recording HRTFs for the sole purpose of sound localization and therefore our recordings only need to capture the implicit direction dependent cues well enough. A schematic overview of HMRR compared to the conventional approach is shown in Figure 2.

3.1. Measurement Approach

Since our method is not actively emitting excitation signals, we have added a third microphone to the recording setup. The microphone is located at a position, that is not affected by echoes of the head and ears. The recordings $x_r(t)$ from this microphone serve as a reference for the source $s(t)$ and we can use it to obtain the HRTFs from the ear recordings with deconvolution by division in frequency domain:

$$H_i(f) = \frac{X_i(f)}{X_r(f)} = \frac{S(f)H_i(f)}{X_r(f)} \tag{8}$$

Equation (8) only holds true in an ideal case where the recording $x_r(t)$ is exactly the source $s(t)$. In a real case the reference recording is affected by the room impulse response (RIR) and can be described as $x_r(t) = s(t) * r_r(t)$. In practice all three recordings are affected by the RIR $r_{pos}(t)$ at the position of the microphone. The deconvolution then yields an approximation of the HRTF:

$$\hat{H}_i(f) = \frac{S(f)H_i(f)R_i(f)}{S(f)R_r(f)} = H_i(f)\frac{R_i(f)}{R_r(f)} \tag{9}$$

This means that our approximation is the real HRTF distorted by the quotient of the room transfer functions at the position of the respective ear $R_i(f)$ and reference microphone $R_r(f)$. The quotient is mainly dependent on the room echoes and we minimize its influence by truncating the calculated impulse responses to a short length. Algazi et al. [1] also used impulse response shortening to eliminate room echoes.

3.2. HMRR Frame Selection Algorithm

Not every source signal is appropriate for approximation of HRTFs with Equation (9). The deconvolution is performed with division of the ear signals by $X_r(f)$. We have to enforce nonzero magnitudes in the spectrum of the reference signal in order to keep the division numerically stable. Additionally, the quality of the approximated HRTFs is dependent on the signal to noise ratio of the source signal. Therefore, we do not use the whole sound signals, but we chop the signals into sound frames

$$\begin{aligned} y_{ij}(t) &= x_i(t + j * f_{size}),\ t \in [0, f_{size}] \\ y_{rj}(t) &= x_r(t + j * f_{size}),\ t \in [0, f_{size}] \end{aligned} \tag{10}$$

where f_{size} is the frame size. We select the sound frame that gives the best deconvolution results by looking at the inverse of $y_{rj}(t)$

$$y_{rj}^{-1}(t) = \mathcal{F}^{-1}\left(\frac{1}{\mathcal{F}(y_{rj}(t))}\right) \tag{11}$$

and convolving it with the original sound frame. In an ideal case the result $d_j(t) = y_{rj}(t) * y_{rj}^{-1}(t)$ of this convolution would be an unit impulse. In practice $d_j(t)$ is an unit impulse with additional noise, because of the mentioned numerical problems and the stationarity assumption of the Discrete Fourier Transform. Nevertheless, we expect the sound frame with the largest peak in $d_j(t)$ to give the best deconvolution results and we find the frame index k by evaluating:

$$k = \arg\max_j |(d_j(f_{size}/2)| \tag{12}$$

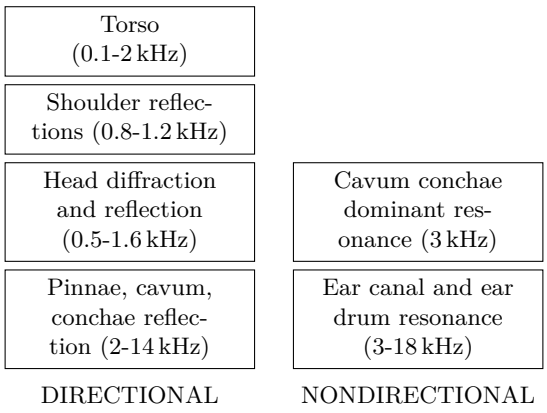

Fig. 3: Impact of different body parts on frequency regions.

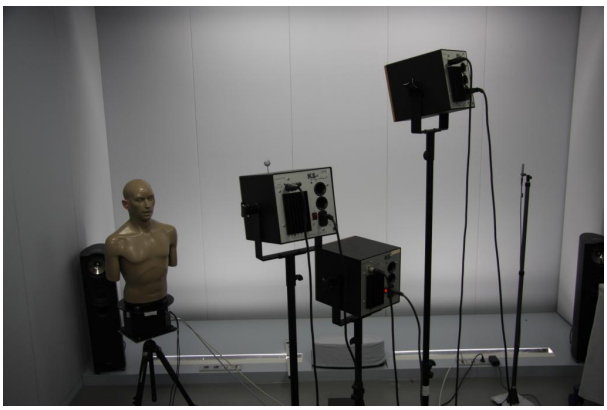

Fig. 4: The experimental setup. The reference microphone is located at the right side of the image.

Due to the convolution operator properties the peak is located at $f_{size}/2$. The frame selection can be improved, if the frequency region of the dominant direction dependent components of the HRTFs is known. As can be seen in Figure 3 for human subjects the region is 0.1-14 kHz and it contains most of the information that is introduced by the subjects body [2]. By band-passing every sound frame $y_{rj}(t)$ before its inversion the frame selection algorithm gives implicitly more weight to the important frequency regions in the frame selection step. The HRTF calculation is performed afterwards with the unfiltered sound frame k.

3.3. Localization with HRTF Approximations

In this Section we briefly discuss the impact of the HRTF approximations and the RIR on the cross-convolution localization algorithm. We will look at the signals in frequency domain and omit the dependency on the frequencies for better readability. Extending Equation (6) with the HRTF approximations and RIRs we get

$$\begin{aligned}
\tilde{S}_{1,\theta,\varphi} &= H_{2,\theta,\varphi}\frac{R_2}{R_r}X_1 = H_{2,\theta,\varphi}\frac{R_2}{R_r}H_{1,\theta_0,\varphi_0}R_{1L}S \\
\tilde{S}_{2,\theta,\varphi} &= H_{1,\theta,\varphi}\frac{R_1}{R_r}X_2 = H_{1,\theta,\varphi}\frac{R_1}{R_r}H_{2,\theta_0,\varphi_0}R_{2L}S \\
\tilde{S}_{1,\theta,\varphi} &\approx \tilde{S}_{2,\theta,\varphi} \iff \theta=\theta_0, \varphi=\varphi_0. \qquad (13)
\end{aligned}$$

Additionally to the RIRs from HMRR in the localization case we can have two additional RIRs R_{1L} and R_{2L}, which correspond to the positions of the ears during the recording of the localization signal. If the robot has not moved between HRTF measurement and localization ($R_i = R_{iL}$), it can be seen from Equation (13) that the cross-correlation criterion will still hold true.

When the robot has moved ($R_i \neq R_{iL}$), this is not obviously the case. But for short impulse response lengths and short ear distances we can assume that the RIRs for both ears will be very similar at each position, meaning $R_1 \approx R_2$ and $R_{1L} \approx R_{2L}$. Therefore, we expect Equation (13) to work even for situations where the robot has moved.

It is also interesting to note, that the RIR of the reference microphone can be completely neglected, since it is a common factor to both ears. This means that the reference microphone can be mounted on the robot or be placed at a random position in the room.

4. EXPERIMENT

We conducted experiments to determine the measurement accuracy of our approach and to study the impact of the RIR on the algorithm in practice. Therefore, we made all recordings in a reverberant environment without special acoustic insulation. Figure 4 shows our experimental setup.

We presented a number of sounds from different directions with a loudspeaker to a KEMAR dummy. The KEMAR itself was mounted on a turntable and could be rotated automatically during the recording

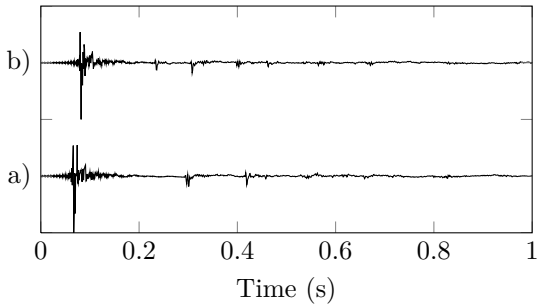

Fig. 5: The RIRs for the two recording positions we used. Position a) is in the middle of the room, b) is close to a wall.

process. The distance between the loudspeaker and the KEMAR was $\geq 1.3\,m$ to ensure far field conditions. The reference microphone was placed at a random position in the room. With this setup we recorded sounds on a spatial grid with three elevations ($\Delta\psi = 10°$) and 144 azimuthal ($\Delta\varphi = 2.5°$) directions at two distinct positions in the room. The Room Impulse Responses for the two positions are shown in Figure 5. Position a) is approximately in the middle of our room and position b) is much closer to a wall. This can be seen by the earlier first reflection at approximately 230 ms which does not exist at position a).

In the first step we created two reference HRTF databases at position a) with the conventional approach, that will serve as ground truth data for the evaluation in Section 5. As excitation signals we used exponential sweeps (20 Hz - 20 kHz, 2 s length) [4, 8] and MLS sequences (2^{17} samples at a sampling rate of 48 kHz) [9]. Additionally, we used these recordings to create HRTFs with HMRR by treating the excitation signal as unknown.

During the second step we recorded a number of sounds from each of the 432 directions at position a):

- 2 music tracks
- 2 different speakers
- 1 random sound signal

As music tracks we selected a pop song with strong vocal parts and a classical track with changing volume and instruments. For the speech signals we recorded two male and one female speaker in a studio environment. A feed from an Internet radio station serves as a random sound source of human speech or music and will be different for every source direction. For the evaluation these signals will serve as excitation signals for HMRR and as test signals for determining the quality of the HRTF databases.

In the third step we relocated the KEMAR to position b) and repeated the measurements from step two. This data will be affected by a different RIR and will serve as additional test signals for studying RIR effects.

In our experiments we collected data for two reference HRTF databases and seven HMRR databases. Additionally, we also recorded five different test signals at two positions in the room.

5. EVALUATION

In this section we discuss our evaluation procedure and results for the HRTF measurements with HMRR. Since we are mostly interested in the question how well the HRTF's spatial cues were captured, we will use localization accuracy as a measure for the quality of the HRTFs. To calculate the accuracy for one database we use the cross-convolution algorithm from Section 2.2 to localize the five test signals from all 432 directions. The percentage of correctly recovered positions is directly related to the quality of the HRTF measurements. For all evaluations we treat angles in a tolerance region of 2.5° as correct. Besides the number of correctly localized positions, we also present the mean angular deviation of the estimated direction to the real one, after correction of front-back confusions.

5.1. Databases from artificial excitation signals

First of all we look at the difference between the conventional approach and HMRR using only artificial excitation signals, because we expected those to give the best results in the HMRR case. Therefore, we created HRTF databases from exponential sweeps and MLS sequences, which we also used to measure the reference databases. Figure 6 shows a comparison of a reference HRIR to one obtained by HMRR with an exponential sweep as the excitation signal. In addition to the difference in SNR the magnitude of the reference HRTF is smoother, especially for higher frequencies. For these frequencies the energy

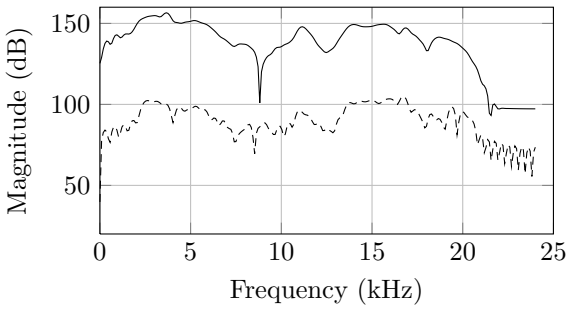

Fig. 6: HRIR ($\psi = 0°$, $\varphi = 0°$) measured with an exponential sweep using the conventional (solid) and HMRR approach (dashed).

Database	accuracy	mean dev.
Reference MLS	85.56%	2.06°
HMRR MLS	61.11%	5.74°
Reference Sweep	86.67%	1.87°
HMRR Sweep	12.78%	54.12°
HMRR Sweep (large)	67.78%	5.56°

Table 1: Localization accuracies for databases created with MLS sequences ($2^{16} - 1$ samples) and exponential sweeps (20 Hz - 22 kHz, 2 s).

of the excitation signal is lower and HMRR is more influenced by this than the conventional approach.

Table 1 shows the accuracy comparison between the reference and HMRR approach for artificial excitation signals. We used both databases to localize all five test signals and the mean of the results is given in the table. With MLS sequences the accuracy measure is 24.3% lower with HMRR than with the conventional approach. But the angular deviation is only slightly worse by 3.68°. This means that most of the additional errors HMRR makes compared to the reference are in a narrow region around the correct solution.

For exponential sweeps and HMRR the accuracy drops to 12.78% with a mean angular deviation of 54.12°, which means that the HRTF measurement is failing to capture the HRTF's spatial cues. The different behavior of HMRR in regards to MLS and exponential sweeps can be explained with the frame selection of the HMRR algorithm. MLS is a broadband signal and all frequencies should get excited even in a short time interval. Exponential sweeps on the other hand are active at only one frequency at a specific time instance. With a small default frame size HMRR is able to find good segments in sparse signals like speech, but a small time window of an exponential sweep is limited to a narrow frequency region. Therefore, HMRR fails to extract a good HRTF from such signals. When setting the frame size to the length of our exponential sweep, the accuracy of the resulting HRTF database rises to 67.78% with a mean angular deviation of 5.56°.

Database	accuracy	mean dev.
Pop music	64.44%	7.89°
Classical music	60.49%	8.11°
Male speaker	63.89%	5.28°
Female speaker	61.67%	4.90°
Random source	63.89%	5.87°

Table 2: Localization accuracies databases created with different natural excitation signals and HMRR.

5.2. Databases from natural excitation signals

The most interesting aspect of HMRR is its use with natural excitation signals. During the second step of our experiment we recorded five sound sources. Now we use each of these sources as the HMRR excitation signal to create five HRTF databases. For each HRTF database we measure the localization accuracy for the remaining four test signals. We have listed the resulting accuracies in Table 2. As before we have given the percentage of correctly localized positions and the mean angular deviation. All databases have an accuracy in the lower 60% range and a mean angular deviation between 5.59° and 8.13°. These results are similar to those we got with artificial excitation signals.

Figures 7 and 8 show the magnitude spectra of the pop song and the female speaker signal. The energy of the music signal is distributed in the frequency region below 18 kHz. The speech signal on the other hand is sparse and most of the energy is in the region below 10 kHz. The similar localization results indicate that HMRR is correctly selecting frames that have enough SNR in the required frequency region regardless of the source characteristics.

The results from the artificial and natural excitations show that the type of the signal is not that important for the deconvolution as long as the algo-

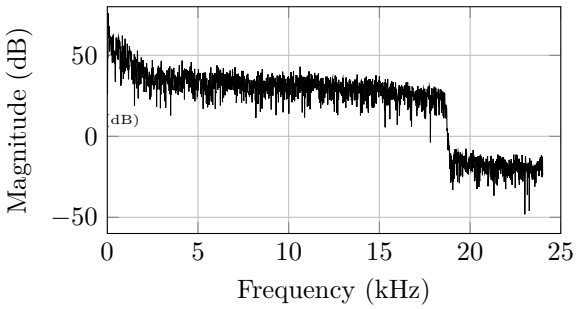

Fig. 7: Magnitude spectrum of the pop music test signal.

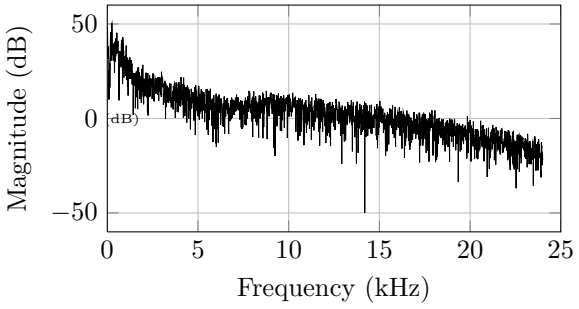

Fig. 8: Magnitude spectrum of the female speaker test signal.

rithm is able to select parts of the signal that have all important frequencies excited.

5.3. RIR dependence of HMRR

As stated in Section 3 HMRR records HRTF approximations which are to some extend dependent on the RIR at the recording position. We claimed that choosing short HRIR lengths decreases this influence and as a default HRIR length we used 128 samples which corresponds to 2.3 ms at a sample rate of 48 kHz. To test if our HMRR databases are dependent on the RIRs we localized test signals recorded at position b) in our experiment with a database that was created with data from position a). Table 3 shows localization accuracies localizing the random source signal for a database created with the pop music signal at position a) with different HRIR lengths. As expected for 128 samples the database is able to localize sounds at both positions in the room with comparable accuracies. For longer impulse responses the accuracy at position a) is getting better

Samples	accuracy at a)	accuracy at b)
128	72.22% 8.75°	66.70% 5.23°
256	80.55% 3.60°	60.23% 5.56°
512	97.22% 2.16°	55.45% 5.60°
1024	99.10% 0.19°	49.79% 5.80°

Table 3: Localization accuracies for databases created at position a) with different HRIR lengths.

as the cross convolution localization algorithm can implicitly use cues from the RIR. At the same time this effect is decreasing the localization performance at position b). This observation confirms our initial assumption about the relationship of the RIR dependence and the length of the HRIR.

6. CONCLUSION

In this paper we have presented a new technique for measuring HRTFs. The HMRR approach does not rely on artificial excitation signals and measures HRTFs by passively recording the environment. Additionally HRTF measurements are not limited to anechoic chambers and can be performed in reverberant environments.

Our experiments have shown that the technique is able to capture the HRTF's spatial cues well and that measured databases can be used to perform sound localization with good accuracy. We studied the effect of the RIR on the measurements and showed that the impact of the RIR on the spatial properties of the HRTFs can be minimized by choosing short impulse response lengths.

ACKNOWLEDGMENT

This work is supported by the DFG excellence initiative research cluster Cognition for Technical Systems CoTeSys (http://www.cotesys.org).

7. REFERENCES

[1] V. Algazi, R. Duda, D. Thompson, and C. Avendano. The CIPIC HRTF database. In *IEEE workshop on applications of signal processing to audio and electroacoustics 2001*, pages 99–102, 2001.

[2] D. Begault. *3-D Sound for Virtual Reality and Multimedia*. Academic Press Professional, 1994.

[3] J. Blauert. *Spatial Hearing*. MIT press Cambridge, Mass., 1974.

[4] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *108th AES Convention*. Audio Engineering Society, 2000.

[5] B. Gardner, K. Martin, et al. HRTF measurements of a KEMAR dummy-head microphone. *MIT Media Lab Perceptual Computing - Technical Report*, 280, 1994.

[6] F. Keyrouz, Y. Naous, and K. Diepold. A new method for binaural 3-D localization based on HRTFs. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, volume 5, 2006.

[7] J. A. MacDonald. An algorithm for the accurate localization of sounds. In *New Directions for Improving Audio Effectiveness*, pages 28–1 – 28–10, Aberdeen Proving Ground, MD 21005 USA, 2005. Army Research Laboratory Human Research and Engineering Directorate.

[8] S. Müller and P. Massarani. Transfer-function measurement with sweeps. *JAES*, 49(6):443–471, June 2001.

[9] D. Rife and J. Vanderkooy. Transfer-function measurement with maximum-length sequences. *J. Audio Eng. Soc*, 37(6):419–444, 1989.

[10] M. Usman, F. Keyrouz, and K. Diepold. Real time humanoid sound source localization and tracking in a highly reverberant environment. In *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, pages 2661–2664, 2008.