HUMANOID BINAURAL HEARING: ADAPTIVE APPROACH

Fakheredine Keyrouz
Notre Dame University - Louaize
Zouk Mosbeh, Lebanon
email: fkeyrouz@ndu.edu.lb

ABSTRACT
When placed in a free sound field, a listener will obstruct an incoming sound wave, and ears, head and body will cause a linear filtering of the sound signal. This filtering is completely and uniquely described by the Head-Related Transfer Function (HRTF). In the general definition of this function all linear properties of the sound transmission are included. All proposed descriptors of localization cues, such as inter-aural differences in arrival-time, in phase, in sound level, as well as monaural cues, are embedded in the HRTF. They can thus be derived from it, whereas the opposite is not generally the case. Motivated by the rich content of the HRTF and by the role of the outer ear to direct, focus, and amplify sound, we present a new binaural method for sound source localization in 3D to be deployed for humanoid binaural hearing. Based on the content of the collected sound signals, the distance between the microphones reconfigures automatically in order to optimize the localization accuracy. Compared to other localization algorithms, the proposed system is outperforming in terms of localization precision and processing power.

KEY WORDS
Binaural hearing, acoustic signal processing, direction of arrival estimation, Bayes procedures, microphones, Head related transfer functions.

1 Introduction

The human hearing organ is a unique signal processor. It comprises acoustic, mechanic, hydro-acoustic, and electric components which, in total, realize a complex receiver and high-resolution spectrum analyzer. Many specialized cells in the auditory pathway contribute to the highly complex signal processing - which by far exceeds the performance of modern computers. The human external ear, pinna, head and torso, transform an incoming sound wave into sound-pressure signals at the two ear drums. The monaural and inter-aural cues resulting from this process, i.e. spectral cues and interaural time and level differences, are employed by the auditory system in the formation of auditory events. All these cues are encapsulated within the HRTF.

It is generally agreed that sound pressures at the ear drum, at the ear-canal entrance, blocked or open, or elsewhere on the center line of the ear canal, have the full spatial information and may be used for binaural recording. Sound source localization techniques using binaural recordings of artificial heads, showed better performances the more human-like these artificial heads are. While several localization techniques using microphone arrays have been proposed, most of them require extensive processing powers and are therefore not suitable for real-time robotic platforms. Very few researchers have dealt with real-time binaural localization utilizing a limited number of microphones for a complete three-dimensional localization.

One sound source localization algorithm based on HRTFs was introduced in [1]. This algorithm is known as the cross-channel algorithm. It simply filters the microphone signal inside the left ear canal with the right HRTF, and the right signal with the left HRTF. This technique was tested for horizontal localization only, using 2 and 4 microphones. In a related work, an enhanced algorithm which combines cross-channel with cross-correlation was proposed [2]. This algorithm computes a coarse azimuth angle based on inter-aural time difference (ITD) using cross-correlation. This coarse angle is utilized by the cross-channel technique in order to enhance its accuracy and diminish its complexity. The presented experimental results are limited to the frontal horizontal plane. We will reproduce this enhanced algorithm in this work and use it as a benchmark for evaluating and comparing our proposed sound locailzation technique.

We have lately introduced a sound source localization technique based on HRTFs [3]. This method was recently enhanced using Bayesian fusion in [4]. The use of Bayesian information fusion considerably increased the localization resolution in a three-dimensional reverberant environment. Compared to existing techniques, the method is able to localize, with higher accuracy, 3D sound sources under high reverberation conditions. The simplicity of the presented algorithm allowed a cost-effective real-time implementation for robotic platforms.

One common and simple method for determining the angle of arrival using pairs of microphones is to estimate the time delay D between two microphone signals x_1 and x_2. This time delay is commonly computed using the generalized cross-correlation function. In order to improve the accuracy of the delay estimate $\hat{D}$, it is desirable to pre-filter x_1 and x_2 prior to cross correlation. The signals x_i are therefore filtered through specified filters H_i yielding the outputs y_i for $i = 1, 2$. Hence, the generalized correlation between x_1 and x_2 is given

DOI: 10.2316/J.2014.216.818-0027

by $\hat{R}^G_{y_1y_2}(\tau) = \int_{-\infty}^{+\infty} W(f)\hat{G}_{x_1x_2}(f)\exp^{j2\pi f\tau}\,df$ where $W(f) = H_1(f)H_2^*(f)$ denotes the general frequency weighting and thhe function $\hat{G}_{x_1x_2}(f)$ is an estimate of the cross power spectrum obtained from finite observations of x_1 and x_2. Hence, the time delay D is evaluated as $D = \underset{\tau}{argmax}\big(\hat{R}^G_{y_1y_2}(\tau)\big)$.

Indeed, depending on the particular form of $W(f)$ and prior information available, it may also be necessary to estimate the generalized weighting. For example, when the role of the pre-filters is to accentuate the signal passed to the correlator at those frequencies at which the coherence or signal-to-noise ratio (SNR) is highest, when $W(f)$ can be expected to be a function of the coherence or signal-and-noise spectra which must either be known or estimated. Table 1 illustrates some common generalized cross correlation weightings The Phase Transform (PHAT) defines a weighting function which is the inverse of the cross power spectral density of the signals. In this technique no individual frequency dominates allowing thus the effects of reverberation to average out. However since it is an inverse of the cross power spectral density, it causes an increase in errors where the signal power is low. On the other hand, the Smooth Coherence Transform (SCOT) defines a weighting function which is the inverse square root of the individual power spectral densities of each received signal. Thereby including contributions from the correlation functions of both left and right signals. The Maximum Likelihood (ML) weighting function minimizes the variance of the time delay estimation. We will use ML function to benchmark and evaluate the performance of our proposed algorithm.

Using the generalized correlation method described above, several binaural models have been put forward to simulate the localization of a sound source in the presence or absence of further, incoherent sound sources, e.g. [5].

In this paper we present a novel adaptive hardware setup which classifies incoming sound signals, adapts the inter-microphone distance, and extracts estimates of HRTFs from incoming signals. Bayesian fusion is then applied to improve the localization precision. The HRTFs of this work are taken from the online CIPIC database. A highly accurate interpolation scheme we have introduced in [6] is then used to obtain a high spatial-resolution database of 28800 HRTFs with one transfer function every 1° in azimuth, covering an elevation from -20° to 60°.

Table 1. Various generalized cross correlation weighting functions.

Window	Weighting functions W(f)	Scope
KORR	1	Direct correlation without a window
SCOT (Smoothed Coherence Transform)	$\frac{1}{G_{x_1x_1}(f)G_{x_2x_2}(f)}$	Suppresses tonal fractions
PHAT (PHAse Transform)	$\frac{1}{\lvert G_{x_1x_2}\rvert}$	Uses only the phase of the cross spectrum
ML (Maximum-Likelihood)	$\frac{\gamma_{12}}{G_{x_1x_2}\lvert 1-\gamma_{12}\rvert}, \gamma_{12} = \frac{G_{x_1x_2}}{\sqrt{G_{x_1x_1}G_{x_2x_2}}}$	Minimizes the variance of the time delay estimation

2 Sound Source Localization System

The proposed sound source localization system deploys four microphones, two placed inside the left and right ear canals, and two outside the ear canals of a Knowles Electronics Mannequin for Acoustic Research (KEMAR). The mannequin is equipped with two silicon outer ears. The outer microphones are allowed to move collinear with respect to the inner ones. The localization system is divided into three main subsystems: right monaural, central binaural, and left monaural. The left and right monaural systems collect the inner and outer signals map them to the frequency domain using Short Time Fourier Transform (STFT) and then divides them to extract the embedded HRTFs. The central binaural system collects both signals inside the ear canals and processes them to determine the corresponding HRTF. These structures are illustrated in Figure 1. The localization system model is shown in Figure 2. More details about the overall system can be found in [4].

3 Audio Signal Classification System

Acoustic signals are classified into five main categories: speech, music, harmonic, non-harmonic and silence [7]. To determine the silence segment, the short-time energy function is used. The incoming acoustic signal entering the ear canal is classified as music if it satisfies one of the following three conditions: its fundamental frequency is small, its Zero Crossing Rate (ZCR) is small, or the variation in its ZCR is small. The acoustic signal is classified as speech if its energy is maximum during a speech and is minimum during a non-speech segment, with the ZCR behaving op-

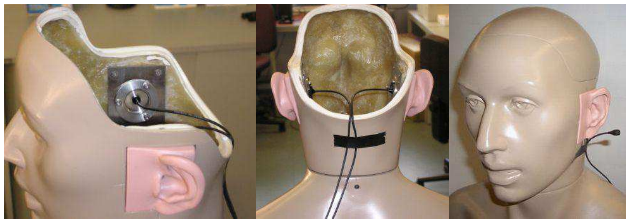

Figure 1. Experimental setup showing the KEMAR head equipped with inner and outer microphones.

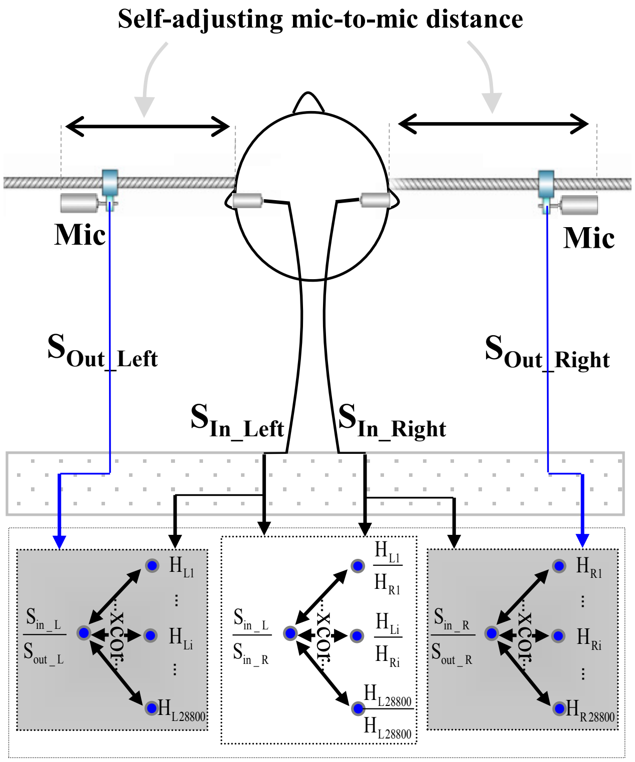

Figure 2. Proposed sound source localization system using 4 microphones.

posite to energy. If harmonic signals cannot be determined as speech or music, they are classified as harmonic. More about signal classification could be found in [8]. Once the audio signals are classified as illustrated in Figure 3, the inner-outer microphone distance is adjusted accordingly in order to maximize the localization efficiency based on the detected signal content.

4 Dynamic Microphone Reconfiguration System

In our older setup, the distance between the inner and outer microphone was fixed [4]. This has limited the localization accuracy of the proposed algorithm because it did not take into consideration the relationship between the frequency content of the microphone signals and the separation between the inner and outer microphones. When the phase delay between the two microphone signals exceeds π, spatial aliasing occurs. This results in a wrong interpretation of the time delays, which in turn results in less accurate localization. Consider a broadband signal, with a maximum frequency f_{max}, incident on a pair of microphones at an angle θ. In order to restrict the phase difference between the pair of microphones to be $\leq \pi$, at the maximum frequency, we let $2\pi f_{max}\tau \leq \pi$, where $\tau = (d/v)sin(\theta)$ is the time delay between the both microphones. The variable d is the distance between microphones, v is the speed of sound. After rearranging these terms, we get

$$d \leq \frac{1}{2}\left(\frac{v}{f_{max}}\right)\frac{1}{sin(\theta)} \tag{1}$$

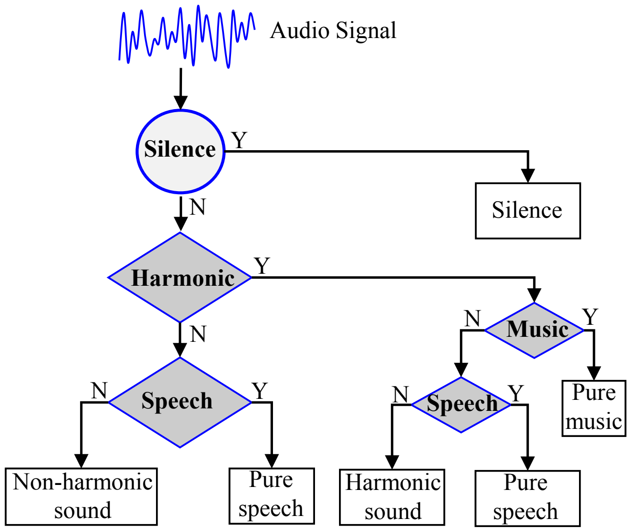

Figure 3. Audio signals classified into five different groups: silence, pure speech, pure music, speech + noise (non-harmonic sound) and speech + music (harmonic sound).

The incident angle of arrival is usually not known and cannot be controlled, we therefore take the worst-case scenario by setting $\theta = 90$. Since the maximum frequency content corresponds to the smallest wavelength present in the signal, we have $f_{max} = \lambda_{min}$. Consequently, $d \leq \lambda_{min}/2$, which means that the distance between the outer and inner microphones should not exceed half the smallest wavelength of the signal. Meeting this condition is a prerequisiste for avoiding spatial aliasing and improving the localization performance. It should be mentioned here, that in the extreme case where two microphones are too close to each others, they will have very small TDOA's. This introduces quantization errors in TDOA computation which in turn results in large localization errors. On the other hand, if two microphones are relatively far from each other, they are less likely to produce a good correlation.

5 Performance Assessment

5.1 Simulation and Experimental Results

We simulate the case where both the left and right outer microphones are allowed to move with respect to the inner ones. Four different signals are used to simulate the classifications done in Figure 3. These signals comprise a pure male speech sound without noise, a pure piano solo sound, a piano sound accompanied with a singer sound, and a female sound corrupted with noise.

Four different simulation tests have been conducted. The first test consisted of the pure male speech signal filtered by 100 different HRTFs. These where randomly chosen from the pool of 28800 different HRTFs. Out of the 100 different locations, 50 were chosen to be having zero elevation. Similarly, the second, third and fourth tests consisted of filtering each of the pure piano, the piano with singer, and the noise-corrupted female sounds, each with 100 different HRTFs.

To simulate the reverberation in our room environment, the image method for room acoustics was used. The simulation setup and room dimensions were defined to match the experimental room environment. A room size of $9.5m \times 7m \times 4m$ was considered. The data received at each microphone was obtained by convolving the broadband source signal with the corresponding transfer functions resulting from the image method between the sources and microphones positions. To simulate real-life environments, high reverberations, i.e. echoes $20dB$ below the signal level, were added to the signal. A sampling frequency of $44.1kHz$ was used.

The proposed algorithm has to first classify the signals, then automatically adjust the microphones to the appropriate mic-to-mic distance in order to satisfy equation 1, and finally run the Bayesian fusion based algorithm for detecting the azimuth and elevation of the now-playing sound signal. Figure 4 shows the mean angular error for different sound signals and mic-to-mic separations.

The average male voice used shows power peaks near 700 Hz in its audio spectrum. For this frequency value the wavelength is about $48cm$. This value bounds the distances between the inner and outer microphones to be less than $24cm$ as dictated by equation 1. Distances above this value, would result in spatial aliasing and therefore increase the localization errors. As illustrated by the blue line (circles) of Figure 4, in the vicinity of $24cm$ the angular error reaches a global minimum of $1.24°$. Below $24cm$ the aliasing error increases slowly. When the distance between the two microphones exceeds $24cm$, the error increases considerably due to font-back confusions.

For the pure piano solo sound used in this work, the highest frequency piano musical note, namely the $C8$ note has a fundamental frequency centered at $4.18kHz$, this value requires the mic-to-mic distance to be approximately less than $4cm$ to avoid spatial aliasing. Above this distance the mean angular error increases. This is illustrated by the green line (squares) of Figure 4.

The sound of a piano accompanied by a male singer shows power peaks around $1.81kHz$. This value corresponds to a maximum mic-to-mic distance of approximately $9.4cm$. Above this distance the error increases as illustrated by the red line (crosses). The female speech corrupted with noise showed power peaks around $480Hz$. This corresponds to a maximum microphone separation of $35cm$. Above this value the error increases as depicted by the black line (stars) in Figure 4.

In our household experimental setup, four different tests have been conducted in order to validate the simulation results. Each of the four different sound signals has been placed at 100 different angles around a KEMAR head and torso in a reverberant room. The level of reverberation in the room was experimentally measured to be $RT = 0.38s$. To keep a fair comparison with the simulation setup, each of the recordings was $350ms$ long. Each microphone was placed $26mm$ inside the ear canal of the dummy head. The experimental results follow the trends of the simulation results, with minor differences due to electronic noise and room geometric mismatches as illustrated by the dashed lines in Figure 4. Both simulation and experimental results showed that the proposed technique is less sensitive to aliasing errors and more affected by separations greater than the optimal distance. This is due to the fact that bigger separations introduce more differences between the outer and inner microphone signals mainly due to reflections, this makes the process of extracting the HRTFs less accurate, and localization more erroneous.

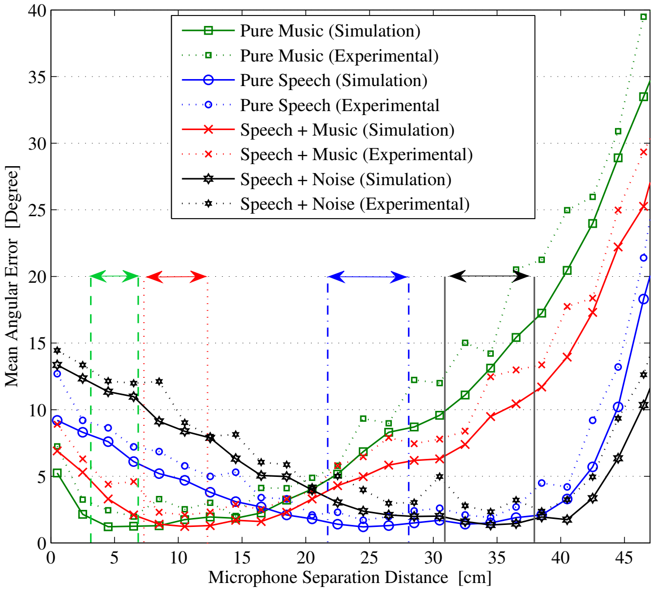

Figure 4. Mean angular error as a function of the distance between the inner and outer microphone. Reverberation time = $0.38s$. SNR = $20dB$.

5.2 Performance Comparison with State-of-the-Art

In this part of our study, two angle of arrival estimation techniques were reproduced and used as a benchmark to our proposed system. These are the ML technique and the enhanced cross-channel technique. For clarity of presentation, experimental results of only the female speech corrupted with noise will be used.

The experimental test consisted of having 100 broadband sound signals filtered with 100 different HRTFs. For each microphone separation distance, ranging from $1cm$

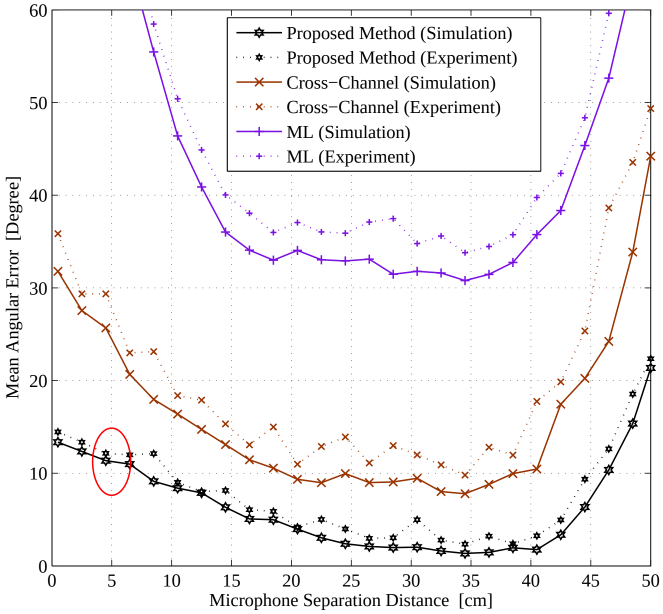

Figure 5. Experimental mean angular error of the proposed method compared to the ML and cross-channel techniques. Reverberation time = $0.38s$. SNR = $20dB$.

till $50cm$, the noisy female speech signal was played over a speaker at 50 different angles in the front hemisphere and 50 angles in the back hemisphere around the humanoid head. Figure 5 illustrates the results. Both simulation and experimental results showed that the proposed method is more tolerant to reverberations and more robust to aliasing errors compared to the cross-channel technique. The combined system with Bayesian fusion and adaptive microphone configuration outperforms all other techniques. This improvement is justified by the fact that the Bayesian fusion method draws intelligence from a conditional probablity table (CPT). This intelligence is not available for the ML and cross-channel techniques. The ellipse in Figure 5 pinpoints the performance of the Bayesian fusion algorithm when the inter-microphone distance is fixed to $5cm$ as in our previous work [4]. A localization accuracy of almost 11° is achieved, by allowing the outer microphone to adjust its position. Similar to the other two methods, the ML technique showed a minimum mean angular error around $35cm$. However, this error is too high compared to the other two methods. This is mainly due to the fact using 4 microphones to localize sound in 3D results in front-back reversal errors which limits the performance of the ML technique considerably.

6 Conclusion

This paper described a robust binaural sound source localization technique based on HRTFs. The novel adaptive hardware setup classifies incoming sound signals, adapts the inter-microphone distance, and extracts estimates of HRTFs from incoming signals. Bayesian fusion is then applied to improve the localization precision. Using only four adaptive microphones, the proposed system showed higher three-dimensional localization accuracy, in the presence of noise and reflections, compared to state-of-the-art methods using the same number of microphones. Based on the presented algorithm, several venues for future work are to be considered. Using wavelet transform instead of STFT, adding echo cancellation to the system and making use of source separation algorithms will make the localization and separation of multiple concurrent sound sources possible.

References

[1] J. A. MacDonald, “A localization algorithm based on head-related transfer functions,” *Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[2] X. Wan and J. Liang, “Robust and low-complexity localization algorithm based on head-related impulse responses and interaural time difference,” *Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 40–46, 2013.

[3] F. Keyrouz, Y. Naous, and K. Diepold, “A new method for binaural 3D localization based on HRTFs,” in *proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006, pp. 341–344.

[4] F. Keyrouz, “Advanced binaural sound localization in 3D for humanoid robots,” *IEEE Transactions on Instrumentation and Measurements*, 2014. (Accepted for publication).

[5] J. Braasch, “Localization in the presence of a distracter and reverberation in the frontal horizontal plane,” *ACUSTICA/acta acustica*, vol. 88, pp. 956–969, 2002.

[6] F. Keyrouz and K. Diepold, “Binaural source localization and spatial audio reproduction for telepresence applications,” *Presence: Teleoperators and Virtual Environments, Special Issue on High Fidelity Telepresence II, Massachussetts Institute of Technology (MIT) Press*, vol. 16, no. 5, pp. 509–522, 2007.

[7] T. Zhang and C. Kuo, “Audio content analysis for online audiovisual data segmentation and classification,” *IEEE Transactions on Speech Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[8] F. Keyrouz, “Automatic self-reconfigurating microphones for humanoid dynamic hearing environments,” in *Proc. IEEE International Symposium on Signal Processing and Information Technology*, Dec 2007, pp. 731–736.