**TECHNISCHE UNIVERSITÄT MÜNCHEN**
Lehrstuhl für Datenverarbeitung

# Efficient Binaural Sound Localization for Humanoid Robots and Telepresence Applications

Fakheredine Keyrouz

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor–Ingenieurs

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing. / Univ. Tokio M. Buss

Prüfer der Dissertation: 1. Univ.-Prof. Dr.-Ing. K. Diepold

2. apl. Prof. Dr.-Ing., Dr.-Ing. habil. H. Fastl

Die Dissertation wurde am 21.02.2008 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 06.06.2008 angenommen.

# Acknowledgments

I would like to thank professor Klaus Diepold for giving me the opportunity to work on my dissertation with him. His support, encouragement, questions, comments, and suggestions have influenced the contents of this work. Furthermore, I thank Professor Patrick Dewilde (Delft University of Technology, Netherlands) for his continuous interest in my research and for his insights and discussions.

In addition, I would like to thank the contribution of Bill Gardner and his colleagues at the Media Laboratory at Massachusetts Institute of Technology for measuring the HRTFs and sharing the database as a courtesy. Besides, I would like to thank Dipl.-Ing. Walter Bamberger for his continuous help throughout my research time. I also thank my colleagues at the data processing institute for the friendly and supportive atmosphere.

Last but not least, I want to thank all my friends and comrades who encouraged me during my work on the thesis. A special and cordial thanks is due to Miss Neunhöffer for the value of proofreading every single chapter of this book. I would also like to thank my parents and family for their unconditioned support. Finally, I want to thank God, for being my constant help.

*And when you have reached the mountain top, then you shall begin to climb.* ***K. G.***

# Contents

# List of Figures

# List of Tables

# Abstract

Auditory signal processing already starts outside the head. The external sound field has to couple into the ear canals. The relative positions of the two ear canals and the sound source lead to a coupling that is strongly dependent on frequency. In this context, not only the two *pinnae* but also the whole head have an important functional role, which is best described as a spatial filtering process. This linear filtering is usually quantified in terms of so-called head-related transfer functions (HRTFs), which can also be interpreted as the directivity characteristics of the ears.

Motivated by the role of the pinnae to direct, focus, and amplify sound, we present a binaural method for localizing sound sources in a three dimensional space to be deployed in telepresence systems. The method is designed to allow robots to localize sound sources using only two microphones placed inside the ear canals of a humanoid head equipped with artificial ears and mounted on a torso. The algorithm relies on extracting important cues of the human binaural auditory system, primarily encapsulated within the HRTF. While existing 3D sound source localization techniques use microphone arrays, the presented method employs two microphones only and is based on a simple correlation mechanism using a generic set of HRTFs.The localization performance is demonstrated through simulation and is further tested in a household environment. While common binaural sound localization methods using only two microphones fail to localize sound accurately in three dimensions without becoming impractically complex, or without using computer vision to augment the acoustic modality, our new localization system demonstrated high precision 3D sound tracking using only two microphones and enabled a low complexity implementation on the humanoid DSP platform.

Based on our new approach, we tackle the challenging task of sound localization in highly reverberant environments as well as the task of sound localization and separation for the underdetermined case where the present sound sources outnumber the available microphones. Simulation and experimental results proved the method to be very noise-tolerant and able to localize sound sources in free space with high precision and low computational complexity, thus suggesting a cost-effective real-time implementation for robotic platforms.

# Chapter 1

# Introduction

## 1.1 Telepresence Framework

Telepresence systems aim at supplying the senses of a human operator with stimuli which are perceptually plausible to an extent that the operator develops a persistent experience of actually being somewhere else, a so-called sense of "presence". The most important stimuli are vision, audio, and haptics. The generic model of telepresence and teleaction is depicted in Figure 1.1. The perceptual world that the operator is experiencing is built up of sensory data provided by a teleoperator, e.g. a tele-robot, located at a remote site. At the local operator site, a human operator is interacting with a multi-modal human-machine interface that renders the sensory data. The human operator manipulates the teleoperator through the interface, which generates the corresponding control signals to be transmitted to the remote site. The integration of audio with other modalities, like vision and haptic not only enhances immersion, but also creates a sense of time-flow within the telepresence operation. In a comparison of auditory and visual perception, Handel [42] arrived at the notion of vision as the generator of the concept of space and the auditory system as a time-keeper. The integration of the auditory and haptic modalities has the potential of mutually enforcing each other [11]. Furthermore, the fact that hearing is an undirected sense is of valuable importance, particularly in case of multiple teleoperators acting at the remote site, since it enables the operator to receive warnings and cues of activities outside his field of view.

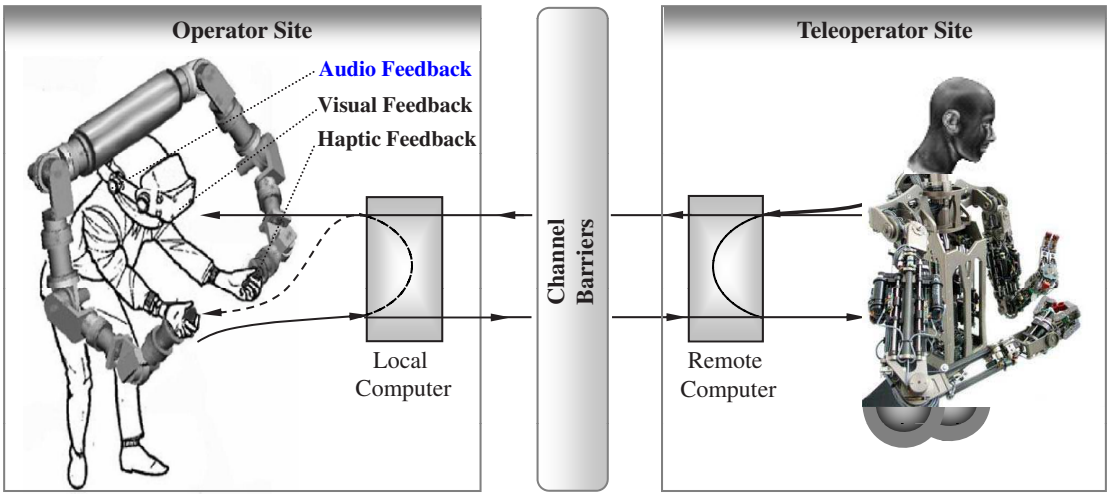


Figure 1.1: A multi-modal telepresence system.

### 1.1.1 Acoustic Telepresence

In a general acoustic teleoperation, we distinguish between two tasks, spatial sound synthesis at the operator site, and binaural source localization at the teleoperator site. In this context, teleoperators are equipped with microphones inserted in their artificial ear canals. These microphones are used to record incoming sound signals which are then analyzed in order to identify the sound source location in the surrounding environment. A typical acoustic telepresence scenario is shown in Figure 1.2. The information about the source location along with a corresponding sound texture is transmitted from the teleoperator site to the operator. There, the incoming sound information along with the direction of arrival information are recombined, using dynamic binaural synthesis of spatial sound, to create an immersive sound impression via headphones. The operator perceives the 3D sound impression of the exact sound source direction as located at the teleoperator site. It should be noted that instead of the original sound texture, it is also possible to use any other natural or synthetic sound texture, which can be positioned in the virtual 3D auditory scene indicated by the direction of arrival information. In the present work, we will focus on the auditory modality at the teleoperator site.

### 1.1.2 Teleoperator Sound Localization

In our acoustic teleoperation, the teleoperator is a humanoid equipped with two small microphones inserted in the ear canals of its artificial head which is mounted

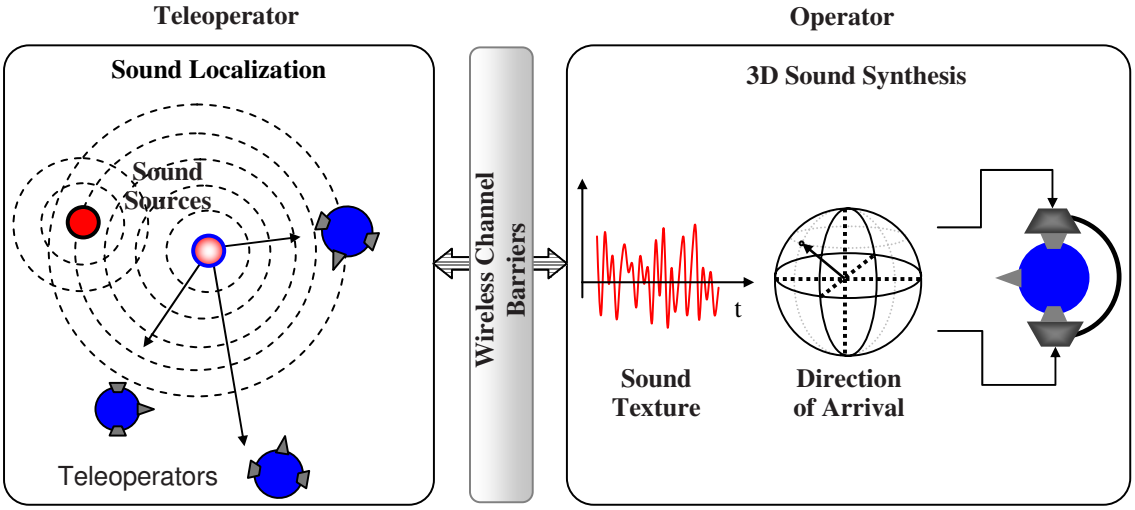


Figure 1.2: Acoustic telepresence scenario.

on a torso. Using two microphones, the teleoperator should localize and track the azimuth and elevation of the sound sources moving in its environment. It is a difficult challenge to use only one pair of microphones on a robot to mimic the hearing capabilities of humans [126]. This task is made even more challenging by the fact that the listening environment is dynamic: sound sources appear, disappear, move and interfere with each other. Until recently, in the fields of robotics and artificial intelligence, the most explored sensory modality is vision. However, unlike eyes (cameras), ears (microphones) do not directly receive spatial information from the surroundings, and rely more on the processing of sensory data to extract auditory cues [53]. Being omnidirectional, human and robot hearing does not require direct line of sight with the sound source. Unlike the human hearing organ, robotic hearing using only two microphones has so far lacked the ability to localize sounds in a three-dimensional environment without becoming impractically complex.

It is known that the incoming sound wave is transformed by the pinnae into sound-pressure signals at the two ear drums. Inspired by the directivity characteristics of the human pinnae and its ability to amplify and focus the sound, this dissertation focuses on building a binaural sound localization system for estimating the three-dimensional position of the sound sources, i.e. the azimuth and elevation angles, while maintaining the low-complexity and real-time implementation requirements for humanoid robots operating in a telepresence environment. A survey of the most frequently used sound localization techniques is presented in the following section.

## 1.2 State-of-the-Art Sound Localization Techniques

The interest in accurate sound localization has rapidly grown in the past few years, mainly due to the fastly increasing necessity of realistic solutions in numerous fields related to audio and acoustics, e.g. 3D sound synthesis, hearing-aid technology, and acoustically-based surveillance and navigation. A large number of localization models have been proposed, most of them are based on microphone arrays, requiring exhaustive processing power in many situations. However, fewer work has dealt with binaural localization where only two microphones are deployed to pinpoint the three dimensional position of a sound, and to allow satisfying real-time localization in acoustically adverse environments.

### 1.2.1 Sound Localization in General

In many everyday listening situations, human beings benefit from having two ears, naturally evolved to analyze concurrent sound sources in various listening environments. For more than a century, research has been conducted to understand which acoustic cues are resolved by the auditory system to localize and separate concurrent sounds. The term *binaural hearing* refers to the mode of functioning of the auditory system of humans or animals using two ears. These ear organs serve as a preprocessor and signal conditioner, they segregate the acoustic cues to help the brain solve tasks related to auditory localization, detection, or recognition.

In humans, the term cocktail-party effect denotes the fact that listeners with healthy binaural hearing capabilities are able to concentrate on one talker in a crowd of concurrent talkers and discriminate the speech of this talker from the rest. Also, binaural hearing is able to suppress noise, reverberance, and sound coloration to a certain extent. 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 differing from the sight of focused visual attention. In many surveillance situations where vision completely fails as the human eyes, or the humanoid cameras, have no direct line of sight with the sound sources, the ability to estimate the direction of the sources of danger relying on the acoustic information becomes of crucial importance.

The process the auditory system undergoes in combining the single cues of the impinging sound waves at the ear drums to a single, or multiple auditory event is not trivial. This holds true, in particular since many psycho-acoustical and neurophysiological details are still unknown, e.g., how the single cues have to be weighted in general. 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 [46].

For the problem of localizing the spatial position of a sound source, a number of models have already been proposed [44]. Most of them are based on using more than two microphones to detect and track sound in a real environment. Mathematical models of sound wave propagation were found to significantly depend on the specific characteristics of the sources and the environment, and are therefore complex and hard to optimize [35]. Adaptive neural network structures have also been proposed to self-adjust a sound localization model to particular environments [104]. These structures disregard the head and pinnae, and create a sort of scanning or beamforming system which can focus on the main source and attenuate reflections as well as other sources. While these networks have been intended to work in specifically controlled milieus, they become very complex in handling multiple sources in reverberant environments. Other methods are designed to mimic the human biological sound localization mechanism by building models of the outer, middle and inner ear, using knowledge of how acoustic events are transduced and transformed by biological auditory systems [41]. Obviously, the difficulty with this approach is that neurophysiologists do not completely understand how living organisms localize sounds.

The human hearing organ is a signal preprocessor stimulating the central nervous system, and providing outstanding signal processing capabilites. It consists of mechanic, acoustic, hydroacoustic, and electric components, which, in total, realize a sensitive receiver and high-resolution spectral analyzer. Binaural hearing does not only have the abilities to focus and discriminate between different sound sources in a host of concurrent sources, but is also able to suppress noise, reverberations, and sound colouration to a certain extent [19].

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, a minimal-complexity sound localization system is needed. Most of the available sound systems today deploy microphone arrays for efficient localization. Those systems which rely on using only two microphones for binaural sound localization are either limited to azimuthal localization only, or they require extensive training sets becoming thus heavily complex and therefore unsuitable for real-time implementation over a robotic platform.

### 1.2.2 Sound Localization Using Two Microphones

The arrival times of the sound wave emitted from a certain source are not exactly the same at the left and right eardrums, due to the different path lengths to both ears as illustrated in Figure 1.3. This arrival-time difference between the left and right ear is called Interaural Time Difference (ITD). The maximal ITD is measured when the sound wave arrives from the side along the axis which intersects both eardrums. In this case, the ITD can be estimated as the distance between the eardrums, $\approx 18$

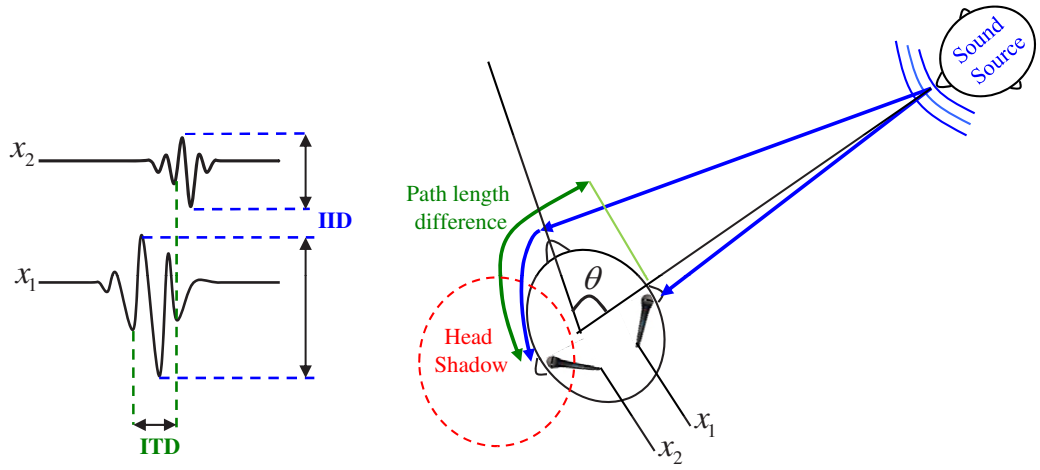


Figure 1.3: Interaural Time/Intensity Difference (ITD/IID).

cm, divided by the speed of sound, $\approx 340$ m/s, to a value of 529 $\mu s$. However, larger ITDs than those are observed in nature. Because of shadowing effects of the head, the measured ITDs can be, depending on the head size, as large as 800 $\mu s$. A model which estimates the ITD in the azimuthal plane for all frequencies throughout the human hearing range has been proposed in [89]. The first physiology-related model for inter-aural time difference computation was proposed by *L. A. Jeffress* back in 1948. The model consists of two delay lines that are connected by several coincidence detectors. A signal arriving at the left ear has to pass the first delay line from left to right. A signal arriving at the right ear travels on the other delay line in the opposite direction. A coincidence detector is activated when it receives simultaneous inputs from both delay lines at the positions that it is connected to. Each of the coincidence detectors is adjusted to a different ITD, due to the limited velocity of propagation of the signals on the delay line. For more details about the *Jeffress* Model, the reader is advised to refer to [54].

The existence of the head between both ears does not only determine the detour the traveling sound wave has to follow, but also causes attenuation of the sound wave

at the contra-lateral eardrum, which leads to Interaural Intensity Differences (IIDs) which are frequently referred to as Interaural Level Differences (ILDs). In contrast to the ITDs, the IIDs are strongly frequency dependent. In the low-frequency range, the human head is small in comparison to the wave length and, therefore, diffraction has only a minor effect on the sound wave. In the high-frequency range, however, the wave length is short as compared to the to the dimensions of the head, and much larger IIDs than in the low-frequency range can be observed. In this frequency region, the IIDs are not only determined by the shape of the head, but are also greatly influenced by the shape of the outer ears. Already in 1877, a geometric model was established to estimate IIDs for various sound-source positions [120].

The Interaural Phase Difference (IPD) refers to the difference in the phase of a wave that reaches each ear, and is dependent on the frequency of the sound wave and on the ITD. For a 1000Hz tone that reaches the left ear 0.5ms before the right. As the wavelength reaches the right ear, it will be 180 degrees out of phase with the wave at the left ear. IPDs are extremely useful as the human ear has the ability to detect differences as small as 3 degrees, and the combination of IPD and ITD, not only aids the listener in determining where the sound stimuli originated from, but also helps identify the frequency of the sound. Once the brain has analyzed IID, ITD, and IID the location of a stationary sound source can be determined with relative accuracy. For fast moving sound sources, however, the human binaural system is slow and less accurate in localization. In this context, the minimum audible movement angle plays an important role in evaluating the localization capability of a given biological or electro-mechanical auditory system. This is defined as the angle through which a sound source has to move in order to be distinguished from a stationary source. A number of investigators have studied this angle subjectively and have reported the maximum ability of humans to follow changes in the location of stimuli over time, i.e., to perceive movements of a sound source [32, 17]. For low rates of movement (15°/s), this angle is about 5°, but as the rate of movement increases, the angle increases progressively to about 21° for a rate of 90°/s [40]. Thus, the binaural system was found to be relatively insensitive to movements at high rates.

One common method for determining the time delay $D$ between the two microphone signals $x_1$ and $x_2$ of Figure 1.3 is the standard cross-correlation function

$$R_{x_1x_2}(\tau) = E[x_1(t)x_2(t-\tau)] \tag{1.1}$$

where $E$ denotes expectation. The argument $\tau$ that maximizes (1.1) provides an estimate of delay. Because of the finite observation time, however, $R_{x_1x_2}(\tau)$ can only be estimated. For example, for ergodic processes, an estimate of the cross correlation is given by

$$\hat{R}_{x_1x_2}(\tau) = \frac{1}{T-\tau}\int_{\tau}^{T} x_1(t)x_2(t-\tau)dt \tag{1.2}$$

where $T$ represents the observation time. In order to improve the accuracy of the delay estimate $\hat{D}$, it is desirable to pre-filter $x_1(t)$ and $x_2(t)$ prior to cross correlation in 1.4. As shown in Figure 1.4, $x_i$ may be filtered through $H_i$ to yield $y_i$ for $i = 1, 2$. The resultant $y_i$ are multiplied, integrated, and squared for a range of time shifts, $\tau$, until the peak is obtained. The time shift causing the peak is an estimate of the true delay $D$. When the filters $H_1(f) = H_2(f) = 1$, $\forall f$, the estimate $\hat{D}$ is imply the abscissa value at which the cross-correlation function peaks. For properly selected filters $H_1(f)$ and $H_2(f)$, the estimation of delay is considerably facilitated. This process is known as the generalized cross correlation [56].

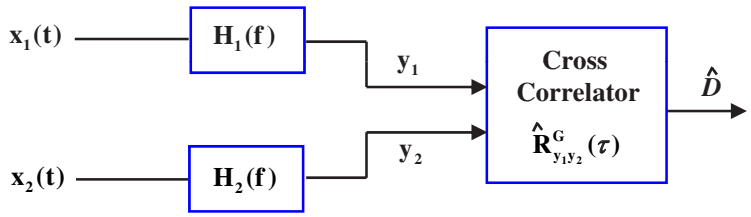


Figure 1.4: Generalized cross-correlation (GCC) process.

The cross correlation between $x_1(t)$ and $x_2(t)$ is related to the cross power spectral density function by a Fourier transform relationship

$$\hat{R}_{x_1x_2}(\tau) = \int_{-\infty}^{+\infty} G_{x_1x_2}(f) \exp^{j2\pi f\tau}\, df \tag{1.3}$$

After $x_1(t)$ and $x_2(t)$ have been filtered as depicted in Figure 1.4, the cross power spectrum between the filter outputs is given by

$$G_{y_1y_2}(f) = H_1(f)H_2^*(f)G_{x_1x_2}(f) \tag{1.4}$$

where $*$ denotes the complex conjugate. Hence, the generalized correlation between $x_1(t)$ and $x_2(t)$ could be written as

$$R^G_{y_1y_2}(\tau) = \int_{-\infty}^{+\infty} W(f)G_{x_1x_2}(f) \exp^{j2\pi f\tau}\, df \tag{1.5}$$

where $W(f) = H_1(f)H_2^*(f)$, denotes the general frequency weighting. In practice, only an estimate $\hat{G}_{x_1x_2}(f)$ of $G_{x_1x_2}$ can be obtained from finite observations of $x_1(t)$ and $x_2(t)$. Consequently, the integral

$$\hat{R}^G_{y_1y_2}(\tau) = \int_{-\infty}^{+\infty} W(f)\hat{G}_{x_1x_2}(f) \exp^{j2\pi f\tau}\, df \tag{1.6}$$

is evaluated and used for estimating the time delay $D$,

$$D = \underset{\tau}{argmax}\big(\hat{R}^{G}_{y_1y_2}(\tau)\big). \tag{1.7}$$

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 ration(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 1.1 illustrates some common generalized cross correlation weightings [88]. 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. Ideally this scheme does not suffer spreading. 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 [64]. We will use these functions to evaluate the performance of our new localization algorithm in chapter 4.

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. [23]. However, a complete model, i.e. a model able to describe all types of binaural phenomena, does not exist yet. This is due to the fact that the human hearing organ is complex

Table 1.1: 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 |

by nature. Sound source separation and localization using only two microphones, has so far lacked the ability to localize sounds in a three-dimensional environment. Methods based on measuring the binaural auditory cues, ITD and IID, resulted in high performance only in the azimuthal plane and except for a slight bias toward the front for sources in the lateral positions. In such methods, the source is assumed to be far enough so that the impinging wavefront is planar. IID and ITD are regarded as frequency dependant; with ITD being more significant at low frequencies while IID is more significant at high frequencies. Reflectors can be placed around the microphones to augment the usual time difference information with intensity difference [90].

From another perspective, average binaural level (ABL), or single-ear spectral energy could be added to the combination of ITD and IID information, to enhance the localization performance [41]. With a pair of microphones, localization is limited to two dimensions, and only up to a front-back ambiguity. Methods based on the robots movement are able to resolve this ambiguity, with high angular acuity ($\pm 2^{\circ}$) [9]. Such methods, however, still lack the ability to localize sound in 3D, and are impractical for sounds of short duration. It is worth noting that such a setup is also subject to mechanical failures due to movement of the robot. The existing algorithms perform poorly in reverberant environments, and techniques trying to compensate for the reverberation by learning a dereverberating filter showed to be very sensitive to even small changes in the acoustic environment [111]. An encouraging and practical method for improving audio source localization by making use of the precedence effect [1] was explored in [137], again adding to the complexity of the system.

Lately, a biologically-based binaural technique based on a probabilistic model was proposed [136]. The technique applies a probabilistic evaluation of a two-dimensional map containing frequency versus time-delay representation of binaural cues, also 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 [61]. The localization model in [29] is based on a neuromorphic microphone taking 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.

A biomimetic algorithm was recently proposed which determines the direction

[1] The precedence effect (PE) describes an illusion produced when two similar sounds are delivered in quick succession (interclick delays of 2-8 msec) from sound sources at different locations so that only a single sound is perceived. The localization of the perceived sound is dominated by the location of the leading sound.

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 weighted against a table of theoretically generated curves in order to determine the direction of arrival of the impinging sound waves. However, due to the symmetrical geometry of the front and back hemispheres, the algorithm applies time-consuming routines to distinguish between the two hemispheres. The algorithm can localize sound source in the horizontal plane only and is limited to the bandwidth of the source and its performance deteriorates in the presence of acoustic and electronic noise [44].

A binaural localization approach based on audio-visual integration was proposed in [107]. The localization method implemented hierarchical integration of visual and auditory processing with hypothetical reasoning on IPD and IID for each subband. In creating hypotheses, the reference data of IPD and IID is calculated by the auditory epipolar geometry on demand. The resulting performance of auditory localization varies according to the relative sound source position. The resolution of the center of the robot is much higher than that of peripherals, indicating similar property of visual fovea (high resolution in the center of human eye). To make the best use of this property, an active direction-pass filter (ADPF), that separates sounds originating from the specified direction by using a pair of microphones, controls the direction of a head by motor movement. In order to recognize sound streams separated by the ADPF, a Hidden Markov Model (HMM) based automatic speech recognition is built with multiple acoustic models trained by the output of the ADPF under different conditions. The method is able to localize sound only in the azimuthal plane and is prone to front/back confusion.

Another method for estimating the location of a sound source using two ears and vision was suggested in [48]. The method is based on extracting localization cues such as ITD, IID, and spectral notches. The authors used a spherical head, having a diameter of 14 cm, and two spiral formed ears of slightly different sizes and with different inclinations to make the extraction of spectral notches possible. Theses notches are supposed to change position linearly as the elevation angle increases. The spiral form of the ears provided a simplified mathematical derivation of the HRTF for the spherical head. The robot was made to learn the HRTF either by supervised learning or by using vision. Audio-motor maps are used to associate sound features to the corresponding location of the sound source and to move the robot to that location. Theses maps are learned using an online vision-based algorithm and are used to provide the appropriate pan and tilt angles for the robot camera. The suggested method has a good accuracy within the possible movements of the head used in the experiments, i.e. -20 to 20 degrees. We will use this method in chapter 4 for comparison purposes with our binaural sound localization technique.

A binaural sound localization method for elevation estimation using a special

pinnae geometry has been suggested in [117]. The pinna has a logarithmic-shaped reflector made of aluminum with a depth of 6 cm and thickness of 0.5 mm. The proposed localization algorithm extracts the spectral cues of the pinnae and estimates an elevation angle accordingly. The spectral cues were identified as those peaks and notches where constructive and destructive interference occur. Based on this approach, sound sources were accurately localized for high elevation angles. for low elevations, however, the performance degraded considerably. The method uses white noise in experimentations and is limited to elevation estimation only.

An artificial human-like dummy head has been introduced in [125] for binaural hearing in telepresence operations. The human-like head is mounted on a humanoid torso and is equipped with artificial ears and two small microphones. The humanoid should operate in a remote environment and its movements are synchronized with the operators head movement. Experimental results show that the listener who is the model for the proposed dummy head can correctly localize sound sources when the head is stationary or synchronized, but that the localization is more precise in the synchronized situation. They also show that in case a listener is not the model of the dummy head, he can still localize in the horizontal plane in the synchronized situation with a good precision. In a further study, the relationship among head shape, head movement, and auditory perception is clarified [124]. In addition, it has been suggested that there is a possibility of building an acoustical telepresence robot with a dummy head of a general shape.

In [108, 106], auditory epipolar geometry was introduced to tackle the sound source localization problem using two microphones. In stereo vision, epipolar geometry [36] is one of the most common methods for extracting depth maps of the observed environment. Depth maps contain information about the azimuth, elevation and distance of objects lying in the field of view of the stereo camera and are used to construct a three dimensional model of the surrounding space. The proposed auditory epipolar geometry in [108] uses ITD and IID cues to compute the azimuth of arrival of sound sources, without providing distance and elevation information. To localize sound sources using two microphones, a set of peaks are first extracted for the left and right channels of the two microphones. Then, the same or similar peaks of the left and right channels are identified as a pair and each pair is used to calculate the ITD and IID cues. Using this piece of information an azimuth of arrival is computed and the sound source is localized. This technique was enhanced in [105] by modeling the humanoid head using the scattering theory in physics [92] to take into consideration the diffraction of sounds around the head for a better approximation of IID and ITD. In this study, the sound source localization module extracts local peaks from the left and right power spectrums and clusters a harmonic sound according to harmonic relationships. Then it calculates ITD and IID of the peaks included in the extracted harmonic sound and calculates distances

between the results and ITD and IID hypotheses created by the scattering theory for each sound direction. The calculated distances are transferred to belief factors on ITD and IID. The belief factors on IID and IPD are integrated to get robust sound localization in the real world. As a result of the integration, a direction with maximum value is regarded as that of the sound source. The resulting system is efficient for localization and extraction of sound at higher frequency and from side directions. However, the system is limited to azimuthal localization in the frontal plane only.

Many models have been put forward to simulate the localization of a sound source in the presence of further, incoherent sound sources, e.g. [23]. For certain conditions these specialized models reach human localization ability, while the localization results of most other computational models are strongly degraded by concurrent sound sources. Another promising approach, [26], achieves robust localization in complex listening scenarios by focusing the analysis of the binaural cues on time instants and frequencies with high inter-aural coherence.

Very recently, we have presented a robotic binaural sound localization method based on hierarchical fuzzy neural networks and a generic set of Head Related Transfer Functions (HRTFs). The robot is a humanoid equipped with the KEMAR[2] artificial head and torso. Inside the ear canals, two small microphones play the role of the eardrums in collecting the impinging sound waves. The incoming signals at both ears are then filtered with a cochlea filter bank built of Dirac deltas. The center frequencies of the filter bank are distributed in a similar way as in the human cochlea [31]. The neural networks are trained using synthesized sound sources placed every 5° in azimuth and elevation, covering elevation angles from −45° to 80°. To improve generalization, the training data was corrupted with noise. Due to fuzzy logic, the method is able to interpolate at its output, locating with high accuracy sound sources at positions which were not used for training, even in presence of strong distortion. In order to achieve high localization accuracy, two different binaural cues are combined, namely, the IIDs and ITDs. As opposed to microphone-array methods, the presented technique uses only two microphones to localize sound sources in a real-time 3D environment [58]. The advantage of this procedure is that often very good results are achieved for stimuli that are similar to the test material. The disadvantages are, however, the long time needed to train the neural network and the fact that the involved processing cannot easily be described analytically. The reader is encouraged to refer to [78], [77], and [71] for more details about the topic. The neural network-based sound localization techniques are beyond the scope of this thesis.

[2]The Knowles Electronics Mannequin for Acoustic Research (KEMAR) is an acoustic research tool which permits reproducible measurements of hearing instrument performance on the head, and of stereophonic sound recordings as heard by human listeners.

### 1.2.3 Microphone Array Based Sound Localization

Getting instantaneous localization of sound in 3D from ITD and IID cues only is a difficult task. For a better performance, today's research focuses on the usage of microphone arrays for a three-dimensional localization under real environmental conditions. While beamforming allows for a localization of multiple sound sources in real time, multiple Kalman filters can be used to exploit temporal information in tracking multiple sources. Such filters with different history lengths can reduce the errors in tracking multiple moving speakers under noisy and echoic environments [103]. Since Kalman filters assume that the state transition is linear, which does not hold in real world, the performance deteriorates severely for moving speakers. For arrays having more than one pair of sensors, two different approaches exist: 1-time difference of arrival (1-TDOA) and 2-time difference of arrival (2-TDOA). While 2-TDOA is a well-studied area and has a low computational complexity, it makes premature decision on an intermediate TDOA, thus discarding useful information. The 1-TDOA approach uses either Phase Transform (PHAT) or maximum likelihood (ML) as the weighting functions. While PHAT works well only when the ambient noise is low, ML works well only when reverberation is small.

In the same context, the Steered Beam (SB) algorithm selects the location in space which maximizes the sum of the delayed received signals. When the SB algorithm is used along with the 1-TDOA, they result in a more robust localization, but their weighting function choices are not well explored yet [116]. An azimuth prediction with accuracy exceeding the one for global extrema detection methods can be achieved by considering the entire cross-correlation waveform [16].

A microphone array of 8 elements using the geometric source separation (GSS) algorithm and a post-filter was studied in [127], resulting in better detection accuracy. Other methods using 6 or more elements per array have been explored [2], [1], [43]. Those methods can be useful in localizing sound in large environments such as airports, but are of small interest in humanoid applications, where simplicity and computational efficiency are crucial.

A robotic spatial sound localization system using 4 microphones arranged on the surface of a spherical robot head was suggested in [49]. The time difference and intensity difference from a sound source to different microphones are analyzed by measuring the HRTFs around the spherical head in an anechoic chamber. While sound arrival time differences were shown to easily be approximated by a theoretical equation, the intensity differences were proven to be more complicated and difficult to be approximated. Hence, only time difference cues were used by the sound localization method at hand. The arrival time differences are calculated from the sound waves of different microphones by the cross-correlation method. By choosing difference microphone pairs, a total of six arrival time differences exists. These differences

are compared with the theoretically pre-calculated arrival time differences which are saved in a database. The distance $d$ between the theoretical and actual arrival time differences is computed. The azimuth and elevation of sound source is then computed as those angles which minimize the error distance $d$. The measurement errors of arrival time difference were shown to become large when the sound source was positioned behind the sphere from the view point of the microphones. Therefore, it was suggested to choose microphone pairs in the front side to the sound source, i.e. to choose the microphone pairs with smaller time difference. We will use this method in chapter 4 for comparison purposes with our localization technique.

A model-based method for sound localization of concurrent and continuous speech sources in a reverberant environment was proposed in [51]. The method applies an algorithm adopted from the echo-avoidance model of the precedence effect to detect the echo-free onsets by specifying a generalized impulse response pattern. An echo-avoidance model assumes that the precedence effect is caused by the neural inhibition of sound localization which depends on the estimated sound-to-echo ratio [50]. This model has two unique properties. First, it uses a generalized pattern of impulse response, which has delay and decay features. Second, the inhibition of sound localization created by this model is a relative one which depends on the sound-to-echo ratio. By using an algorithm adopted from the echo-avoidance model, echoes can be estimated as the maximum effects of their preceding sound by the generalized pattern of impulse response, regardless of the type of sound and the condition of environment. Echo-free onsets can be detected as onsets which have high estimated ratios of sound to echo. The advantages of the model-based onset detection are its flexibility with respect to sound level and its insensitivity to nontransient noise. This method detects onsets after band-pass filtering. Hence, the overlap of a sound component with different sound components of other sound sources is significantly decreased. Three microphones arranged in a triangular form were available for sound localization. Fine structure time differences were calculated from the zero-crossing points between different microphone pairs. They were integrated into an azimuth histogram by the restrictions between them. Two sound sources were localized in both an anechoic chamber and a normal reverberant room. The time segment needed for localization was 0.52 s and the accuracy was a few degrees in both environments. We shall use this method in chapter 5 for comparison purposes with our concurrent sound localization technique.

An algorithm for multiple moving speaker tracking using a microphone array of 8 sensors installed on a mobile robot was introduced in [102]. The localization is based on time delay of arrival estimation, and multiple Kalman filters. The time delay estimation localizes multiple sound sources based on beamforming in real time. Non-linear movements are tracked using a set of Kalman filters with different history lengths in order to reduce errors in tracking multiple moving speakers under noisy

and echoic environments. For quantitative evaluation of the tracking, motion references of sound sources and a mobile robot were measured accurately by ultrasonic 3D tag sensors. The system tracked three simultaneous sound sources in a room exhibiting large reverberation. We will use this method in chapter 5 for comparison purposes with our localization technique.

An accurate sound localization method using 8 microphones and applying the simple TDOA algorithm to localize sound sources in three dimensions was introduced in [128]. Using cross-correlation, this method determines the delay between the signals captured by the different microphones. A major limitation of this approach is that the correlation is strongly dependent on the statistical properties of the source signal. Since most signals, including voice, are generally low-pass, the correlation between adjacent samples is high and generates cross-correlation peaks that can be very wide. The problem of wide cross-correlation peaks is solved by whitening the spectrum of the signals prior to computing the cross-correlation. However, after applying the whitened cross-correlation method, each frequency bin of the spectrum is forced to contribute the same amount to the final correlation, even if the signal at that frequency is dominated by noise. This makes the system less robust to noise, while making detection of voice (which has a narrow bandwidth) more difficult. In order to counter the problem, a weighting function of the spectrum was introduced which gives more weight to regions in the spectrum where the local signal-to-noise ratio (SNR) is the highest. The overall system is able to perform localization even on short-duration sounds and does not require the use of any noise cancelation method. The precision of the localization is $3^\circ$ over 3 meters. We will use this method in chapter 6 for comparison purposes with our localization technique.

As an alternative to microphone arrays, much psychoacoustic research has been performed on human beings and animals to isolate the individual cues of sound localization [30]. One novel approach is sound localization based on HRTFs, which are also called anatomical transfer functions (ATFs). These functions describe the filtering of sound on its way to the inner ear. The HRTFs will be the main focus of the next chapter.

## 1.3 Main Contributions and Overview

### 1.3.1 Main Contributions

We have addressed the challenging task of binaural sound localization using a pair of microphones inserted in the ear canals of a humanoid head mounted on a torso and operating in a general telepresence environment. This task is made even more challenging by the fact that the listening environment is dynamic: sound sources

appear, disappear, move and interfere with each other. The today existing sound localization models fail to localize sound in both azimuth and elevation using only two microphones. Common binaural methods are limited to localizing sound sources in either azimuth or elevation. Those methods which try to achieve three dimensional localization, using only two sensors, become impractically complex by relying heavily on training sets for every environment, or by using computer vision to augment the acoustic modality. Hence, successful sound localization techniques are based on microphone arrays to detect and track sound in a real environment. Microphone arrays require exhaustive processing powers and are therefore not suited for our simple hardware setup deploying two artificial ears.

We have proposed a unifying framework for novel three-dimensional sound localization methods to be implemented on a humanoid robot. The initial proposal is based on dividing the ear signals with the left and right HRTFs and subsequently taking the maximum correlation coefficient as a pointer to the source location. this method is enhanced using proper state-space HRTF inversion. In addition, a new algorithm called cross convolution was developed to further decrease the computational requirements of the initial method. Nevertheless, with the help of a simple properly tuned Kalman filter, a ROI was introduced to account for fast moving sound sources. The efficiency of the new algorithm suggests a cost-effective implementation for robot platforms and allows fast localization of moving sound sources.

Using the presented methods, we have addressed the challenging task of binaural concurrent sound source localization and separation in reverberant environments. We presented a new algorithm for binaural localization of concurrent sound sources in both azimuth and elevation. Compared to existing techniques using microphone arrays for the same purpose, our algorithm is less complex and very accurate. Beside localization, we have proposed a sound source separation algorithm which proved to be outperforming compared to other blind source separation methods solving the same determined problem under the same conditions.

For highly reverberant environments, a new algorithm using four microphones is presented. Bayesian information fusion is then used to increase the localization resolution in a three-dimensional reverberant environment. Compared to existing techniques, the method is able to localize sound sources in three dimensions, under high reverberation conditions, with fewer sensors and higher accuracy.

### 1.3.2 Overview and Organization of the Thesis

**Chapter 2: Binaural Techniques**

In chapter 2, we summarize all important aspects of HRTFs, ranging from their measurement procedure, their time and frequency domain visualization, to their

recent deployment in our sound localization system. In section 2.7, we introduce three methods for HRTF order reduction, namely the Balanced Model Truncation (BMT), Diffuse-Field Equalization (DFE), and the Principal Component Analysis (PCA). These techniques are used in the following chapters to help reducing the localization processing time of a HRTF-based sound localization system. In contrast to conventional methods utilizing HRTFs to synthesize sound for virtual reality, we use the HRTFs for real-life sound detection.

**Chapter 3: Binaural Sound Source Localization Based on HRTFs**

A novel sound localization technique based on HRTFs and matched filtering is proposed in section 3.1 of this chapter. As opposed to sound localization methods based on microphones arrays, the method localizes sound in azimuth and elevation by using only two small microphones placed inside the ear canal of a humanoid head mounted on a torso.

In section 3.2 two novel techniques for accurate HRTF interpolation and robust inversion are presented. In section 3.2.2, an accurate HRTF state-space interpolation technique is introduced to ensure the availability of sufficient HRTFs for higher precision localization performance. In section 3.3, a stable inverse of the HRTFs is computed using state-space inversion based on inner-outer factorization. This inversion technique is deployed in chapter 4 to stabilize the localization of a novel humanoid binaural sound system.

**Chapter 4: Enhanced Sound Source Localization Techniques**

In this chapter, we have considerably improved the initial matched filtering approach to sound localization using the reduced and inverted HRTFs, as in chapters 2 and 3. This set up proved to be able to localize sound sources up to a very high precision in free space. In addition, a cross convolution algorithm for real time localization and tracking is proposed. Applying HRTFs for sound localization together with extended Kalman filtering, we are able to accurately track moving sound sources in real time in a highly reverberant environment. This algorithm uses only two microphones and requires no prior knowledge of the sound signals.

**Chapter 5: Concurrent Sound Source Localization and Separation**

In chapter 5, we combine blind source separation and binaural localization for tracking concurrent sound sources using only two microphones. Section 5.2 describes an algorithm for two sound sources that iteratively adapts the coefficients of a Multiple Input Multiple Output (MIMO) system and provides the two statistically independent source signals. This well-known separation method exploiting the non-stationarity of the sources is used to retrieve two speakers from two convolutive mixtures in real-time. By using a simple relation between blind source separation and system identification, the HRTFs that filtered the sound sources can be de-

termined under the condition of an anechoic environment. The second algorithm, presented in section 5.3, applies Short-Time Fourier Transform (STFT) to the ear signals and makes use of the sparseness of the sources in a time-frequency domain. As in each frequency band the normalized time-frequency patterns of speech signals cluster around the HRTF values, the interaural HRTFs can be retrieved. The positions of the sources are finally determined by a database lookup. With the respective HRTFs of the database, the sources can be separated by inversion of the HRTF-system in case of two concurrent sound sources or by L1-norm minimization in case of more than two sources.

**Chapter 6: Sound Localization in Highly Reverberant Environments**

In this chapter, a novel monaural 3D sound localization technique is presented. The proposed system, an upgrade of monaural-based localization techniques, uses two microphones: one inserted within the ear canal of a humanoid head equipped with an artificial ear, and the second held outside the ear. The outer microphone is small enough to avoid reflections that might contribute to localization errors. The system exploits the spectral information of the signals from the two microphones in such a way that a simple correlation mechanism, using a generic set of HRTFs, is used to localize the sound sources. The main focus of this chapter is on the detection of sound events under severe acoustic conditions, i.e. high reverberation and background noise. The location of the sound source obtained from monaural and binaural observation sensors is fused using a properly tuned Bayesian network in order to increase the localization resolution in a three-dimensional reverberant environment.

# Chapter 2

# Binaural Techniques

## 2.1 The Head Related Transfer Function

Within the framework of human sound localization, it is generally accepted that the head pinnae modify the spectra of incoming sounds in a way that depends on the angle of incidence of the sound relative to the head. The spectral changes produced by the head and pinna can be used to estimate the localization of a sound source. This has been confirmed by measurements in the ear canal of human observers and by measurements using realistic models of the human head [19], [133]. 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 is called the HRTF or equivalently, the head related impulse response (HRIR). The HRTFs capture the diffraction of sound waves by the human or humanoid torso, shoulders, head, and outer ears and hence vary in a complex way with azimuth, elevation and frequency. In addition, the HRTFs depend on the morphology of the listener's body, and therefore vary significantly from person to person. The HRTF for a particular individual is called his or her individualized HRTF.

Binaural sound reproduction builds on the concept that our auditory percepts are primarily formed on the basis of only two inputs, specifically the sound pressure signals formed at our two eardrums. If these are recorded using small microphones inserted inside the ears of listeners, and reproduced authentically when played back, then all acoustic cues and all spatial aspects are accessible to the listeners for producing authentic replicas of the original auditory percepts.

The main application area for HRTFs is the reproduction of binaural hearing for virtual reality application. In this context, instead of being picked up inside the ears of the listener, the signals to the two ears, called binaural signals, are generated electronically by the use of filters representing HRTFs. In this case the method will be denoted binaural synthesis. In binaural synthesis, a virtual sound source is created by simply convolving a sound signal with a pair of HRTFs [67]. The success of binaural synthesis strongly depends on details of the procedures applied for determining and realizing HRTFs, such as physical aspects of the measurement situation, post-processing of data, and implementation as digital filters. If the full spatial information is maintained in the resulting binaural signals, we say that the HRTFs contain all properties of the sound transmission and all descriptors of localization cues.

## 2.2 HRTF Cues

The acoustic cues for sound localization have been studied for over a century [19]. The most reliable cues used in the localization of sounds, some of them depending upon a comparison of the signals reaching the two ears, could be generally divided into five main categories: 1) ITD, 2) IID/ILD, 3) monaural cues, 4) head rotation and 5) Interaural Coherence (IC). While ITDs and IIDs are the most commonly used for modeling sound localization systems, monaural cues, resulting from the complex topographic shape of the pinnae, play a very important role in decoding the elevation information of the impinging sound sources. Head rotation is commonly used to resolve front/back ambiguity as well as the cone of confusion problem[1]. The IC cues are mainly used by acousticians to extract the auditory spatial properties of a certain environment [21]. The limits to the auditory system localization ability are determined by the limits of detecting and analyzing the above mentioned cues.

The HRTF can be interpreted as the directivity characteristics of the two ears and shows a complex pattern of peaks and dips which varies systematically with the direction of the sound source relative to the head and which is unique for each three dimensional direction. Recently, a method to robustly extract the frequencies of the pinna spectral notches from the measured HRIR, distinguishing them from other confounding features, has been properly devised [112]. Scientific study of HRTFs has focused on two questions: what acoustic cues do people use to localize sound sources, and what parts of the body are responsible for generating those cues. This

[1]For a given interaural spectrum, there exists a surface or a locus of points corresponding to sound source locations for which ITDs and ILDs are identical. A cone (whose axis is a line drawn between the ears and whose origin is the center of the head) is a fair approximation of this surface. Using only ITDs/ILDs there are no cues available to resolve positional ambiguity on such a cone.

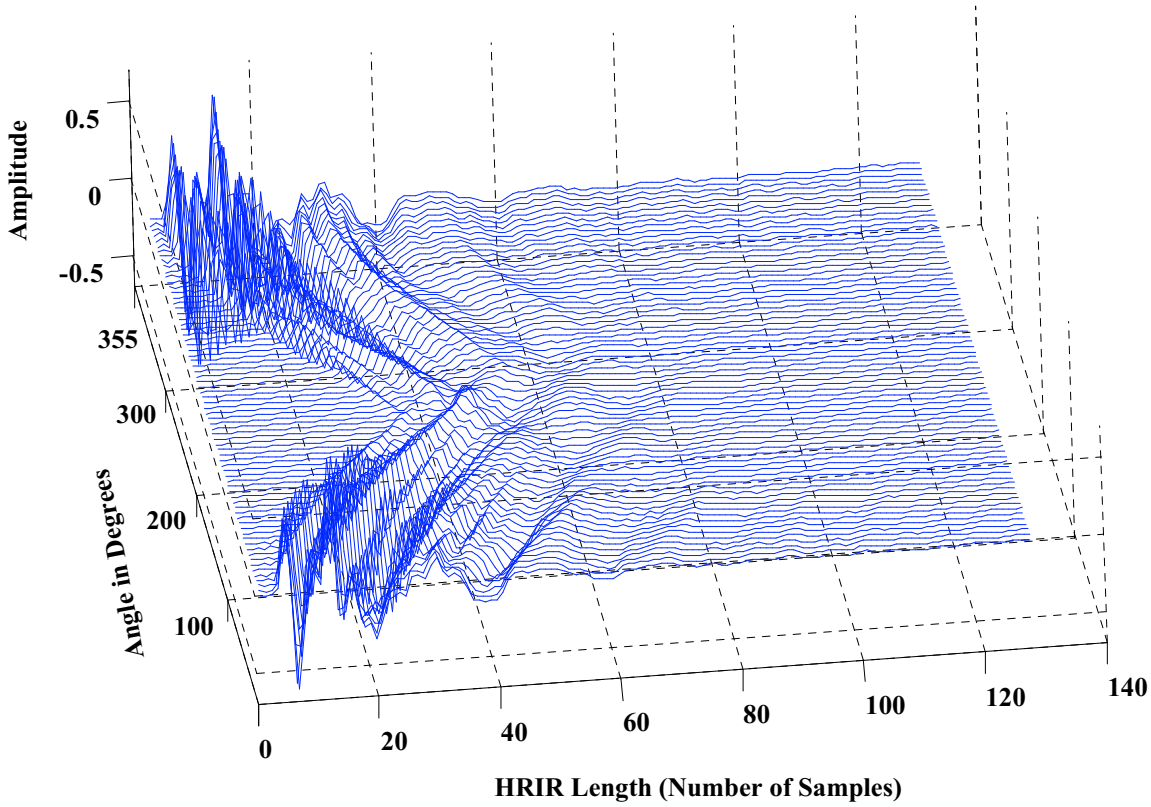


Figure 2.1: Measured right-ear Head Related Impulse Responses (HRIRs) for source locations in the horizontal plane (Elevation = 0°).

research is presented systematically in [19], and several useful reviews are available [98, 100].

## 2.3 HRTF Properties

Because size and shape of the external ear vary greatly between persons, the task of relating the anthropometry of the pinna to the localization cues it creates is cumbersome. It remains unclear how macroscopic features of HRTFs correspond to perceptually meaningful directional cues. It stays difficult to exactly pinpoint which peaks and notches in HRTF magnitude frequency responses correspond to the specific azimuths and elevations of the sound source locations.

Many methods have been trying to uncover structure in HRTF data by visually

comparing subsets of HRTFs sharing the same azimuth, elevation, and frequency in the time or frequency domain. It was shown in [27] how the location of a spectral notch near 7 kHz changes as a function of elevation. In [4], the authors show how diffraction effects due to the head and shoulder can be seen as secondary echoes in time domain versions of HRTFs. Within this framework, it was suggested in [28] how peaks in plots of spatial location vs. HRTF for a fixed frequency could correspond to perceptually preferred directions in space. The method in [5] shows that a composite model combining independent contributions of the pinna and of the head and torso considered as a unit results in a good HRTF approximation. This model isolates the pinna by treating it as if it were mounted on an infinite plane, and leads to a significant simplification in the pinna response, the so-called Pinna-Related Transfer Function (PRTF). In the following, we shall distinguish between two visualizations of the HRTF, the time domain impulse response and the frequency domain transfer function.

### 2.3.1 Time Domain Impulse Responses

The measurements shown in Figure 2.1 are taken for the KEMAR right ear and are plotted as a function of azimuth in the horizontal plane (elevation = 0°). Looking at this figure, one can observe the relatively large amplitude of the initial peaks in the impulse responses corresponding to azimuth +90°, i.e. to the location where the source is directly facing the right ear. While the source is moving towards the left ear, the HRIR peaks fade down slowly due to increased head shadowing, and reach a minimum at the contralateral location where the source is directly facing the left ear.

Figure 2.2 shows measured HRIRs as a function of elevation in the median plane (azimuth 0°) and in the vertical plane corresponding to azimuth 90°. One can also see elevation-related effects as there is a slight difference in arrival times for positive and negative elevations. From Figures 2.1 and 2.2, we can observe that, in addition to the initial peak, the measured HRIRs contain many secondary peaks. This is caused by the numerous spatial cues as well as by the complex morphological structure of the outer ear, although these effects could also be related to the inherent noise in the measurement process.

Figure 2.3 is the image representation of the KEMAR's HRIRs. The figure shows the responses of the left and right ear to an impulsive source in the horizontal and median plane. The strength of the response is represented using different levels of brightness. Looking at the left ear data, we can see that when the sound source is moving in the horizontal plane, the strongest response is reached at an azimuth angle of 270° or equivalently -90°, the weakest and mostly delayed response occurs

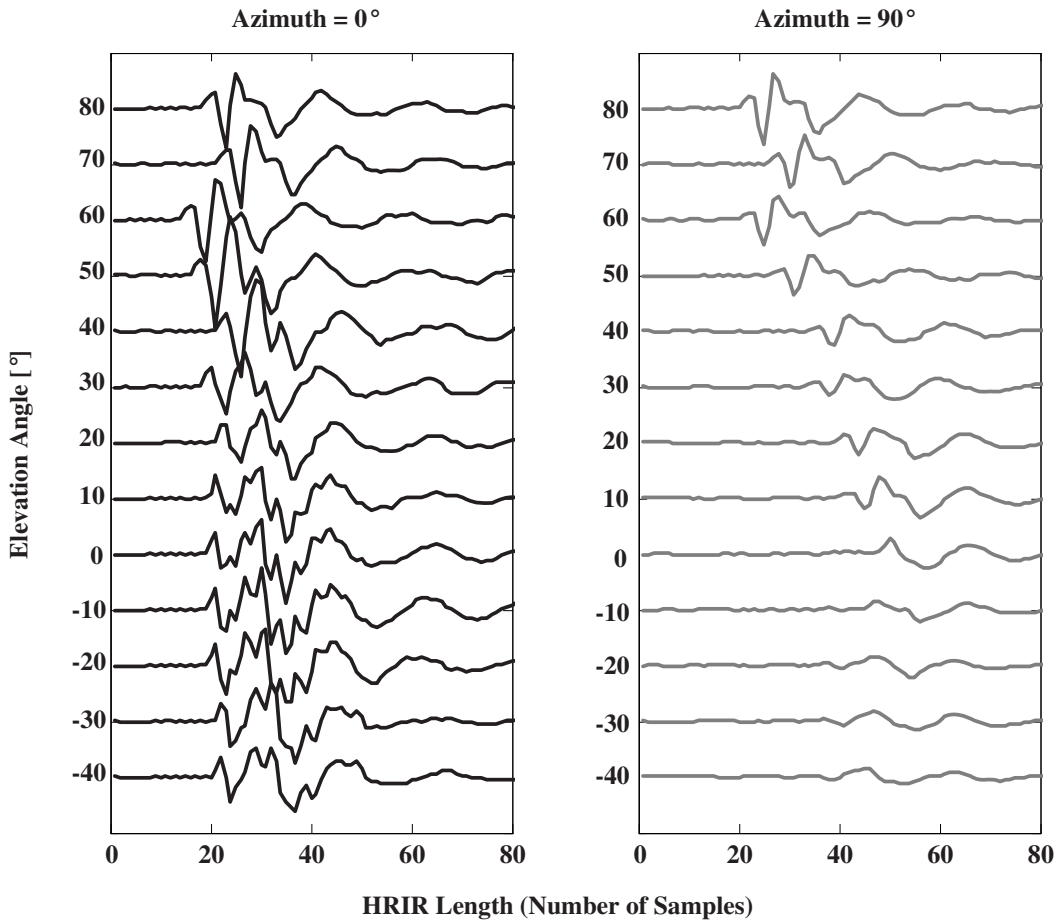


Figure 2.2: Measured right-ear Head Related Impulse Responses (HRIRs) for source locations in the two vertical planes: Azimuth 0° and Azimuth 90°.

at azimuth = 90°. The right ear response shows a directly opposite behavior. In a nutshell, we observe a pronounced variation of the impulse response in the horizontal plane as a function of azimuth. However, this is not the case for the median plane.

When the sound moves around the head in the median plane, as shown in the right subplots of Figure 2.3, it reaches the left and right ears at almost the same time. There is no apparent difference for the strength of the impulse response among all of the elevation angles, and the arrival time is more or less the same. The main changes are in the relative arrival times and strengths of the pinna reflections. This explains why people have trouble distinguishing front from back when sounds are located in the median plane. This phenomenon is well-known as the front/back confusion problem and people often resolve it by head motion.

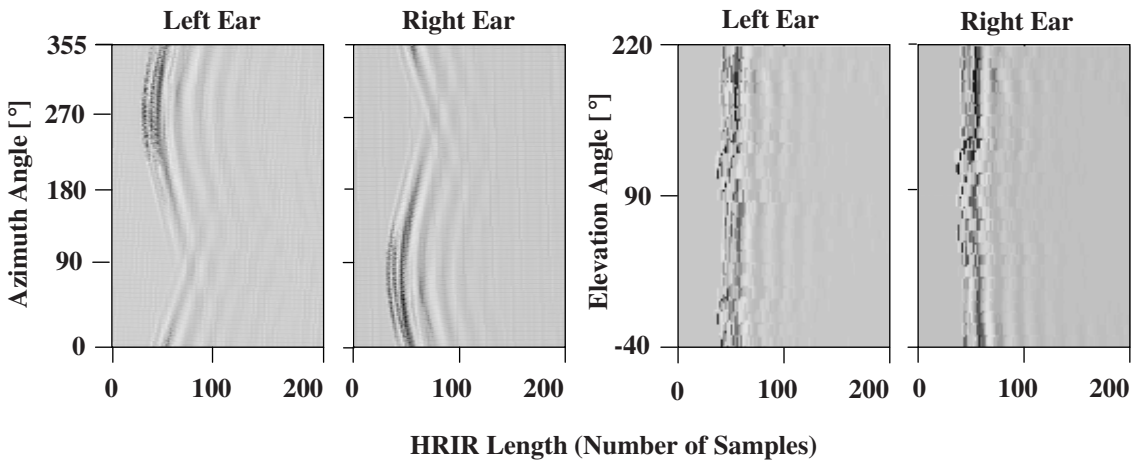


Figure 2.3: Left two plots: HRIRs in the horizontal plane (elevation=0°). Right two plots: HRIRs in the median plane (azimuth=0°).

### 2.3.2 Frequency Domain Transfer Functions

In the frequency domain, the HRTFs exhibit a complex pattern. Several secondary peaks and notches show up in the magnitude spectra, as depicted in Figure 2.4. These features are due to the filtering of the pinna, head, and torso. The visualization of the HRTFs in the frequency domain allows us to distinguish between two structural effects: diffraction effects due to the head and elevation effects due to the pinna.

For some frequencies and incident angles, the head has an amplifying effect on an incident plane wave impinging on it at certain points. This magnification is due to diffraction. There are some locations on the contralateral side of the head where this effect occurs, even though the head directly blocks or shadows the contralateral ear.

Diffraction effects in the left ear HRTFs are highlighted in Figure 2.4. In the right subplot, the HRTFs corresponding to azimuths 80° to 65° contain a low frequency main lobe that attains its greatest width at azimuths 80°. This main lobe is representative of an amplification effect the head has on lower frequencies due to diffraction on the contralateral side, i.e. on the side where the sound source is totally shadowed by the head. High-frequency amplification effects can also be seen in the ipsilateral HRTFs, i.e. the HRTFs corresponding to those positions where the sound source is directly facing the ear. This effect is mainly due to reflections caused by the outer ear's proximity to the head. The amplification regions are pinpointed

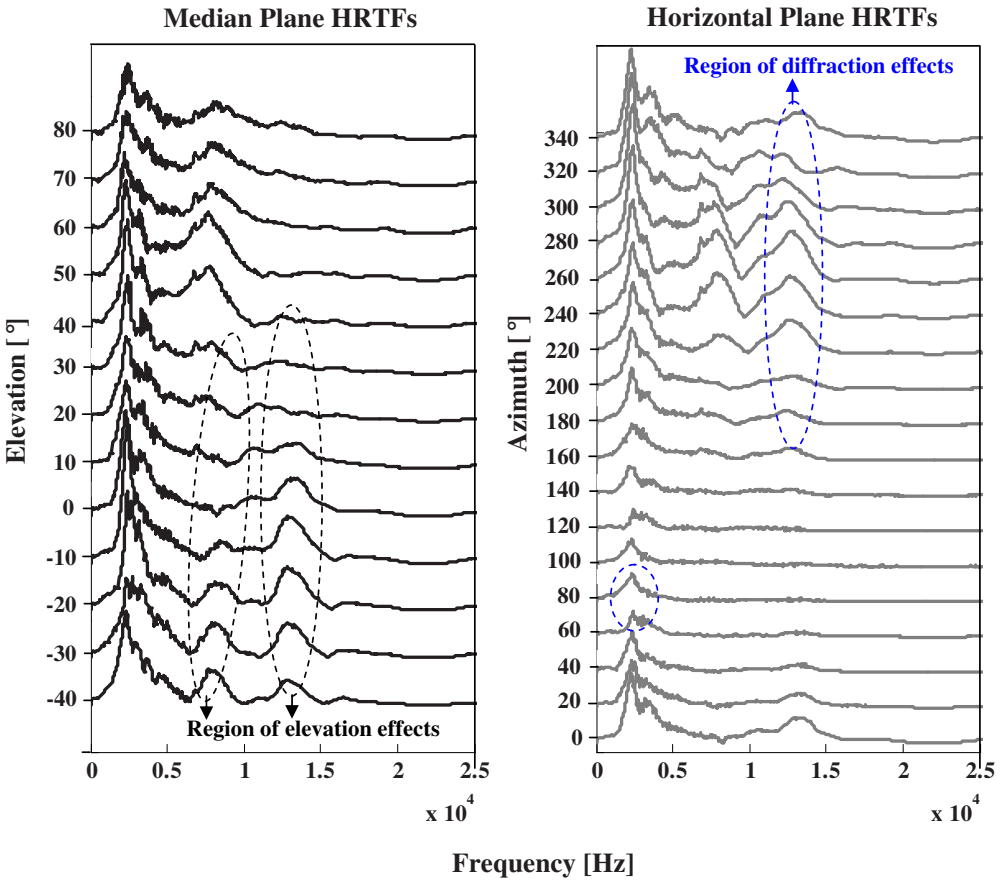


Figure 2.4: Head-Related Transfer Functions (HRTFs). Left: variations in the median plane (azimuth = 0° ). Right: variations in the horizontal plane (elevation = 0° ).

in Figure 2.4.

The HRTF spectral shaping related to elevation is believed to be directly associated to the external ear [34]. Hence, frequencies around 6-8 kHz ($\lambda = 4.1 - 5.5$cm) play an essential role in the elevation decoding process, since these frequencies have wavelengths similar to the anthropomorphic lengths of the pinna, and therefore strongly interact with the pinna. As seen in Figure 2.4, noticeable patterns appear around these frequencies which have been proved psychoacoustically to be associated to the sound elevation perception [19].

Further elevation effects can be seen in Figure 2.4. At low frequencies, a first main notch appears around 7 kHz and slowly migrates upwards in frequency as elevation increases. At high frequencies, a narrow peak appears around 12 kHz for lower elevations in the median plane, and gradually flattens out for higher elevations.

Several scientific studies have tried to disclose the relationship between peaks, notches and spectral shapes of certain HRTF databases, as well as their systematical contribution to the perception of azimuth and elevation. Consequently, several frequency-domain based signal processing techniques attempted to parameterize or compute HRTFs models. For example, frequency-domain HRTF interpolation, as well as frequency based PCA and pole-zero modeling have been repeatedly studied [45].

## 2.4 HRTF Measurements

Until today, almost all available HRTF data has been obtained from direct acoustic measurements [114, 47, 130, 134, 39]. A conventional HRTF measurement procedure consists of placing an artificial humanoid head and torso on a motorized turntable which can be rotated accurately to any required azimuth. A speaker mounted on a boom stand enables accurate positioning of the speaker to any elevation with respect to the humanoid. Thus, the measurements can be made one elevation at a time, by setting the speaker to the proper elevation and then rotating the humanoid to each azimuth. Such a conventional HRTF measurement procedure is tedious and require expensive specialized equipment. Normally, only a sparse spatial grid of HRTFs is measured.

Several HRTF databases have been made publicly available [122, 93, 91, 6]. In addition, commercial products are available for measuring individualized HRTFs [13]. Although various methods for speeding up the measurement process have been proposed [141, 38], the acoustic measurement of accurate, high-resolution HRTFs remains a time-consuming process that requires special equipment that is not widely available. This presents a serious obstacle to the widespread usage of HRTFs. Hence, as an alternative to the measurement process, numerous investigations of signal processing methods for approximating and modeling the HRTFs have been conducted [118]. For example, physical HRTF modeling techniques allowed individualized HRTFs to be reproduced by adjusting the parameters of such models in order to fit the anthropometry of the pinnae, head, and torso for a certain individual.

## 2.5 First Binaural Localization Model

The celebrated duplex theory proposed [113] is known to be the very first binaural hearing model. The theory is based on the idea that humans use the ITD at low frequencies and the ILD at high frequencies to judge the lateral (azimuthal) angle, the angle between a ray to the sound source and the median plane. These difference

cues have the important property that they are independent of the source spectrum. Moreover, by approximating the head by a rigid sphere, Rayleigh gave a quantitative explanation of how the ITD and ILD are produced by diffraction and scattering of the incident waves by the listener's head.

Although the spherical-head model is obviously only a first approximation to a real head, it continues to be a source of insight into HRTF behavior [4, 119, 135]. Rayleigh understood that interaural difference cues vanish on the median plane, and thus cannot resolve front/back confusion. In general, the ITD and ILD constrain the ray to the source to lie on a cone called the "cone of confusion". It is common experience that when we are not sure if a source is in front of us or in our back, we turn our heads, which works well unless the sounds are very brief. In a study of the effects of head motion, it was speculated that to resolve front/back confusion for brief sounds, the auditory system somehow uses the *selective sound shadow* of the outer ears or pinnae. For longer duration sounds, it was shown that head motion cues dominate other cues [131]. The role of the pinnae in resolving the front/back confusion problem was subsequently confirmed by other researchers [37, 110]. All culminated to the undisputed fact that the pinnae are perceptually important and have a major effect on the HRTF.

## 2.6 HRTFs and Sound Localization

In our telepresence scenario, a robot placed at a remote site is supposed to detect the location, in terms of azimuth and elevation, of the sound sources randomly partitioned in a certain environment. This angle information along with the sound signals are sent through the wireless channel to the remote human operator site, where sound synthesis using HRTFs takes place for 3D virtual reproduction of the robots auditory space.

To date, the usage of HRTFs basically revolved around binaural sound synthesis, e.g. surround sound by headphones, 3D auditory displays, binaural mixing consoles. As opposed to virtual reality sound synthesis using HRTFs, the present thesis aims at using the HRTFs for real-life robotic sound localization. Our target is to build a robotic sound source localizer using the cues encapsulated within generic HRTF measurements, and then use these measurements and develop a low-complexity model for azimuth and elevation estimation. Towards this end, we interpolate, truncate, and invert the HRTFs. Interpolation is used to avoid the complex and time-consuming measurement process. Truncation of the HRTF database is needed for faster signal processing, and stable inversion ensures better sound localization results.

Experiments have shown that measured individualized HRTFs can undergo a great deal of distortion (i.e. smoothing, reduction, etc.) and still be relatively effective at generating spatialized sound [19]. This implies that the reduced HRTF still contains 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, thus shortening the length of each HRTF and consequently reducing the overall localization processing time.

The research group at MIT Media Lab has made extensive measurements using KEMAR. They provide a data set consisting of 710 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. Note that audio professionals consider the length of 512 samples for the measured HRTFs as rather short. When it is necessary to use longer filters the computational burden increases accordingly. Therefore, 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. The original HRTFs containing the 512 coefficients of the FIR filter will be denoted as $H_{512}^{FIR}$. Using the reduced dataset, we present a novel approach in chapter 4 to localize sound sources using only two microphones in a real environment.

## 2.7 HRTFs Reduction Techniques

### 2.7.1 Diffuse-Field Equalization

In order to reduce the computational burden for convolution, we aim at shortening the length of the FIR filter representation of the originally measured HRTFs. This needs to be done while preserving the main characteristics of the measured impulse responses. We adopt the algorithm proposed by [101] for a Diffuse-Field Equalization (DFE). In DFE, a reference spectrum is derived by computing a power-average over all HRTFs for each ear and taking the square root of this average spectrum. Diffuse-field equalized HRTFs are obtained by de-convolving the original transfer function by the diffuse-field reference HRTF of that ear. This leads to the result that the factors that are not dependent on the incident-angle, such as the ear canal resonance, are removed. The DFE is achieved according to a four step procedure:

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, e.g. ear canal resonance, loudspeaker and microphone responses [101].

3. Smooth the magnitude response using a critical-band auditory smoothing technique [95].

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 DFE HRTF database is denoted as $H_{128}^{FIR}$. Figure 2.5 shows one example of a diffuse-field equalized HRTF filter response $H_{128}^{FIR}$ in comparison to the originally measured HRTF $H_{512}^{FIR}$. The reduced HRTF follows the general trend of the original one with some deviation at high frequencies.

To quantify the accuracy of the DFE process, spectral signal-to-error power ratios (SER) have been computed for the difference between both models. For the 710 impulse modeled impulse responses, the SER values were in the range of 20-37 dB with an average of 30 dB.

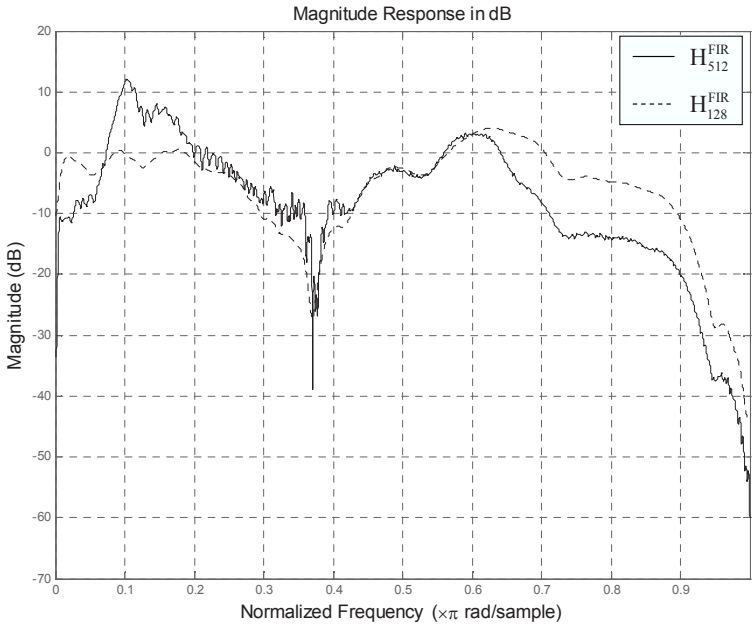


Figure 2.5: Magnitude response of the original 512-FIR (solid) and the reduced 128-FIR (dashed) of an HRTF (left ear, 0° azimuth)

### 2.7.2 Balanced Model Truncation

In order to examine to which extent the HRTF can be reduced while still preserving the characteristic information making 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 $H_{128}^{FIR}$. A detailed description of the BMT technique is given in [15]. However, a brief outline will be presented here. For applying BMT, we determine a linear time-invariant state-space system, which realizes the filter. We start using the 128-coefficient FIR filter $H_{128}^{FIR}$. The transfer function of this filter can be written as: $F(z) = c_0 + c_1 \cdot z + c_2 \cdot z^2 + \ldots + c_n z^n$, where $n = 127$. Note that we follow the notation of positive exponents for the z-transform. This filter can be represented as state-space difference equations:

$$\begin{aligned} x(k+1) &= A \cdot x(k) + B \cdot u(k) \\ y(k) &= C \cdot x(k) + D \cdot u(k) \end{aligned} \tag{2.1}$$

Then, a transformation matrix $T$ 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 $(n - m)$. Figure 2.6 shows the BMT-reduced IIR filter representation ($m = 25$) in comparison to the FIR ($n = 128$) for one example of a HRTF. The transfer function of the IIR filter follows the general trend of the FIR filter with small deviation at high frequencies. Quantitative SER ratios have been computed for the difference between the FIR and IIR models. For all the 710 impulse responses that we modeled, SERs were in the range of 24-36 dB, with an average of 29 dB.

### 2.7.3 Principal Component Analysis

In order to examine to which extent the HRTF can be further reduced while still preserving the characteristic information which makes it unique, we reduce the previously derived diffused-field HRTF data set, $H_{128}^{FIR}$, by applying a Principal Component Analysis (PCA).

PCA has already been applied to HRTFs [96, 99, 87]. All the applications pointed out substantial data reduction, as this method allows the description of all HRTF data with merely 4-7 basic functions and their corresponding weights. The

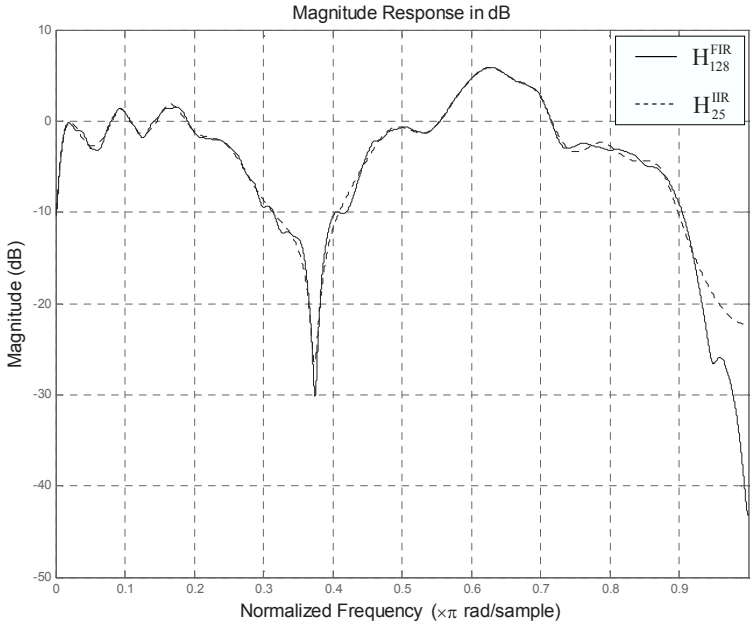


Figure 2.6: Magnitude response of the 128-FIR (solid line) and the reduced 25-IIR (dashed line) of a HRTF (left ear, 0° azimuth).

PCA of personalized HRTFs efficiently pointed out some common properties and some main differences in HRTFs for various test subjects.

The first step in PCA is the computation of a $p \times p$ covariance matrix $S_k$, which corresponds to the $k_{th}$ HRTF $H_k$, with $k = 1, 2, ..., 710$. The entries of this matrix are given by

$$S_{i,j} = \frac{1}{N} \sum_{k} H_{k,i} \cdot H_{k,j}, \quad i, j = 1, 2, ..., p, \tag{2.2}$$

where $p$ is the total number of frequency samples (512 in this case), and $H_{k,i}$ is the magnitude of the $k$th HRTF at the $i$th frequency. The $S_k$ matrix provides a measure of similarity across the HRTFs for each pair of frequencies.

A basis-function matrix $BF_k$ is then derived from the eigenvectors of the covariance matrix $S_k$. This matrix contains $q$ basis functions. The basis functions are chosen to be those eigenvectors of $S_k$ which correspond to the $q$ largest eigenvalues.

$$BF_k = \begin{bmatrix} EV_1 & EV_2 & \cdots & EV_q \end{bmatrix} \tag{2.3}$$

where $BF_k$ is $p \times q$, and $EV$ denotes an eigenvector. The HRTF can then be modeled as a linear combination of several weighted basis functions. For a given HRTF, the

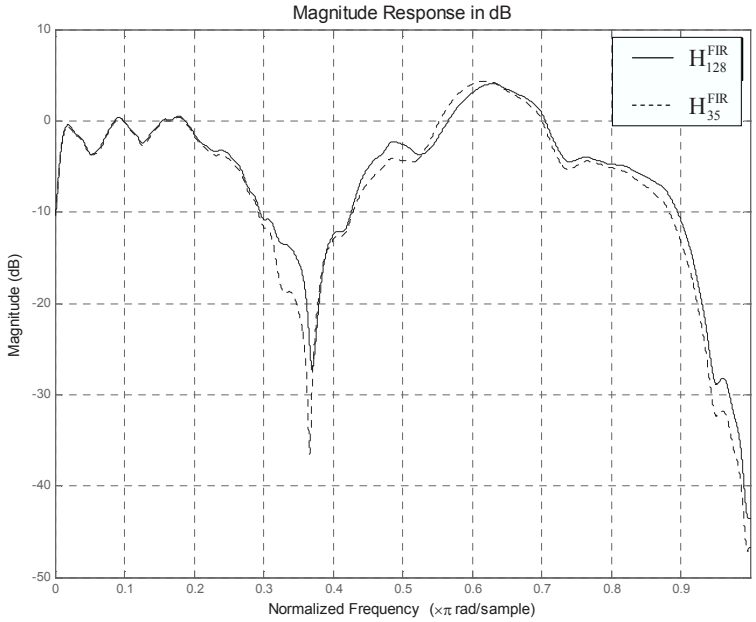


Figure 2.7: Magnitude response of the 128-FIR (solid line) and the reduced 35-FIR (dashed line) of a HRTF (left ear, 5° azimuth).

weights representing the contribution of each basis function to that HRTF are given by

$$w_k(\theta, \phi) = BF_k^T \times H_k(\theta, \phi) \quad k = 1, 2, ..., 710, \tag{2.4}$$

where $w_k$ is $q \times 1$, and $H_k$ is $p \times 1$. Note that if the terms are rearranged, the HRTF magnitude vector is equal to a weighted sum of the basis vectors:

$$H_k(\theta, \phi) = BF_k \times w_k(\theta, \phi) \quad k = 1, 2, ..., 710. \tag{2.5}$$

However, this equality holds if and only if $q = p$, or the maximum possible number of eigenvectors and basis vectors is retained. In practice, $q << p$, here we take $q = 5, 10, 15, 20, 25, 50$. We found that 91.337% of the total variance in the data is captured by taking the first 20 basis functions only. Taking more basis functions can further reduce the error between the measured and approximated HRTFs, however more calculation time and storing space are required. Once the reduced basis functions are selected, the weighting matrix is calculated. These two matrices (correlation matrix and weighting matrix) are stored e.g. in DSP memory, for a fast computation of reduced HRTFs. A thorough description of the PCA technique in modeling HRTFs is available in [87]. 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 data set. Figure 2.7 plots the difference between both

FIR models. The SER values were found to fall in the range of 23-37 dB, with an average of 30 dB.

In the following chapter, we will introduce two novel HRTF inversion and interpolation techniques which help stabilizing and increasing the localization accuracy of our new sound localization system.

# Chapter 3

# Binaural Sound Source Localization Based on HRTFs

In this chapter, we propose a new binaural sound source localization technique based on using only two small microphones placed inside the ear canal of a robot's head. The head is equipped with artificial ears and is mounted on a torso. In contrast to existing sound source localization methods, we employ a matched filtering approach using the HRTFs applied to the signals collected by the two microphones. This set-up proves to be able to localize sound sources in free space with high precision. Note that, so far, HRTFs have mainly been used for synthesis of spatial sound, while we are using them here for sound source localization.

Furthermore, this chapter introduces two novel approaches for inverting and spatially interpolating the HRTFs to be used later in our sound localization algorithm. Using HRTF data, we create proper matrix transfer functions or equivalently appropriate state-space realizations. We base our method on the factorization of a block Loewner matrix into a product of generalized observability and controllability matrices. We recollect certain properties to be satisfied by the Loewner matrix, and use it to construct a minimal state-space realization of an interpolating matrix transfer function.

## 3.1 A Novel Approach To Sound Localization

The new algorithm for binaural sound source localization relies on a simple correlation approach [82]. We assume that we receive a signal originating from a sound

source located at a certain position. The signal is recorded using a left and a right microphone, both microphones being placed inside the canals of artificial ears mounted on a manikin. The signal received by each ear can be modeled as the original signal filtered with the HRTF corresponding to the given ear and to the specific direction of the sound source.

If the two received signals (left and right) were to be filtered with the inverse of the correct HRTFs, then both output signals should be identical to the original mono signal of the sound source. However, the system does not have information about the sound source position. Nevertheless, the result of filtering the signal received by the left ear with the correct inverse left $H_{\mathrm{L}}$ should be identical to the signal received by the right ear filtered by the correct inverse right $H_{\mathrm{R}}$.

In order to determine the direction from which the sound is arriving, the two signals must be filtered by the inverse of all HRTFs. The pair of inverse HRTFs that produces a pair of filtered signals resembling each other the most should correspond to the direction of the sound source. The resemblance of the filtered signal pair is determined using a simple correlation function. The direction of the sound source is assumed to correspond to the HRTF pair with the highest correlation. Therefore, we base our localization on the obtained maximum for the correlation factor $c$. Moreover, to insure an accurate localization decision, the minimum distance measure, $d$, 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 flow of this algorithm is shown in Table 3.1.

Table 3.1: Initial Sound localization algorithm.

| |
|---|
| $S_L(t)$ = Received signal inside left ear<br>$S_R(t)$ = Received signal inside right ear<br>$n$ = Number of HRTFs |
| for $i=1:n$<br>$x_L^{(i)}(t) = S_L(t) \otimes H^{-1}{}_L^{(i)}$<br>$x_R^{(i)}(t) = S_R(t) \otimes H^{-1}{}_R^{(i)}$<br>$c^{(i)} = corr(x_L^{(i)}(t),\ x_R^{(i)}(t))$<br>$d^{(i)} = \sum_m (x_L^{(i)}(t) - x_R^{(i)}(t))^2$<br>end |
| $\otimes$ = Convolution<br>$m$ = HRTF length |

The filtering in time domain yields a significant computational complexity. Hence, the algorithm is applied in the frequency domain to allow the use of fast correlation and convolution techniques. The signals along with the reduced HRTFs are converted into the frequency domain using the Fast Fourier Transform (FFT). Doing this, all the filtering and correlation operations are changed to simple array multiplications and divisions [63]. This method is illustrated in Figure 3.1.

We assume that the reduced HRTFs have been computed before using DFE and PCA and the results are stored in memory. Once a block of 128 audio samples are recorded by the left and right microphones, each of them is transformed into frequency domain using an FFT of length 128. Subsequently, 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 of length 128, and 710 correlation operations necessary. After all the correlations are computed, the maximum correlation value is taken to provide the direction from which the sound is arriving. The block-oriented processing of 128 samples recorded at a sampling frequency of 44.1kHz produces a minimum processing delay of about 22$\mu$sec, which is well below the allowable processing delay.

The above-mentioned sound localization technique depends on finding the inverse filters of every HRTF, and saving it for later use in localization. The inverse filter was directly made available by simply exchanging the values of the numerator and denominator using FFT. However, the inverted HRTF filters obtained using the FFT method are unstable, especially that all HRTFs include a linear-phase component, i.e. pure delay, which is vital for maintaining the correct inter-aural time difference.

A very efficient method which handles this problem, and ensures stability, by simply translating the unstable inverse into an anti-causal yet bounded inverse, is the state-space inversion method we will introduce in the following section. We have used this inversion technique to considerably stabilize our sound localizer [72].

## 3.2 Efficient State-Space HRTF Interpolation

The efficiency of our sound localization approach is directly dependent on the quality and availability of the HRTFs. The quality of the HRTFs strongly depends on details of the procedures applied for realizing HRTFs, such as physical aspects of the measurement process, post-processing of data, and the implementation as digital filters. Since these procedures are complex, time-consuming, and require expensive specialized equipment, only a discrete grid of spatially sampled HRTFs is available.

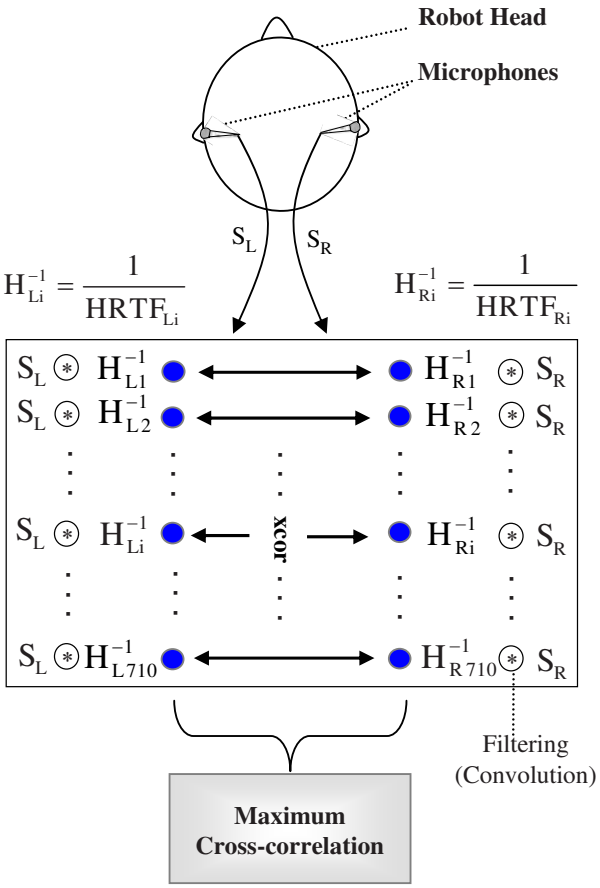


Figure 3.1: Flowchart of the sound localization algorithm.

Spatial interpolation techniques can be employed to increase the resolution of the grid by computing HRTFs corresponding to 3D positions which are located between the recorded functions.

The lowest directional resolution for sampling HRTFs in order to ensure that interpolations between them do not introduce audible errors has recently been properly devised [109].

Our aim is to construct a rational interpolation method which, given two neighboring angles, could correctly interpolate a good number of HRTFs in between. Towards this end, we recollect some of the existing interpolation techniques, and use them to study the performance of the rational interpolation presented here.

### 3.2.1 Previous Interpolation Methods

The bilinear method [14] is a simple and direct way used to perform HRTF interpolation. It consists of computing the binaural response corresponding to a given point on the horizontal circle as weighted mean of the measured binaural responses and associated with the two adjacent points to the desired point. The discrete Fourier Transform (DFT) [97] is used to interpolate binaural impulse responses in the time domain. The method inserts all HRTFs column-wise in one common $H$ matrix. It then computes the DFT of every row and appends it with zeros before applying inverse DFT to obtain an oversampled matrix, where the oversampled columns correspond to the interpolated HRTFs. The plenacoustic function [121] and the mathematical spline function [115] have also been used to interpolate HRTFs.

In [97], the authors investigated the effect of arrival time correction (initial time-delay equalization) on the accuracy of the three interpolation methods, namely, bilinear interpolation, discrete Fourier transform and third-order spline function. The initial time-delay equalization was demonstrated to increase the accuracy of all proposed interpolation techniques. In this chapter, we use the outcome of their experiments to compare and evaluate the performance of the rational HRTF interpolation we are presenting. Before we present the framework in which we applied rational interpolation, we will have a brief review of the scalar and matrix rational interpolation methods.

### 3.2.2 Formulation of the Rational Interpolation Problem

The rational interpolation problem was first solved for the scalar transfer function case [10]. Due to its frequent occurrences in linear system theory, a transfer-function matrix solution rather than the scalar representation was basically of more relevance, hence a state-space description of the problem was later devised in [8]. We will recapitulate some of the key properties of the scalar rational interpolation problem, before tackling the state-space rational interpolation problem.

#### Scalar Rational Interpolation

Consider the array of points $P := \{(x_i, y_i), i = 1, \ldots, N\}$, with $x_i \neq x_j$, and $x_i, y_i \in \mathbb{C}$. The fundamental rational interpolation problem [8] is to parameterize all rational functions $y(x) = \frac{b(x)}{a(x)}$ having minimal Smith-McMillan degree, which interpolate the above points. If $x_i \neq x_j$ for $i \neq j$, then the desired rational function must satisfy $y(x_i) = y_i$, for $i = 1, \ldots, n$. For this purpose, the rational interpolating function

$y(x)$ could be selected such that

$$\sum_{i=1}^{n} c_i \frac{y(x) - y_i}{x - x_i} = 0, \quad c_i \neq 0. \tag{3.1}$$

The function $y(x)$ is the desired interpolation function, for which we clearly have $y(x_i) = y_i$, if $c_i \neq 0$. The goal is to minimize the degree of the adopted $y$. One way to do this is to consider a summation as in (3.1) containing only $q < n$ summands, for any set of non-zero coefficients $c_i$, then the rational function $y$, of degree $q - 1$, interpolates the first $q$ points. Making use of the freedom in selecting the $c_i$, we then try to interpolate the remaining $n - q$ points. Let $\mathbf{c} := [c_1 \dots c_q]^T$; in order for the remaining $n - q$ points to be interpolated, $\mathbf{c}$ must satisfy

$$\sum_{i=1}^{q} c_i \frac{y_{q+j} - y_i}{x_{q+j} - x_i} = 0, \quad i = 1, 2, \dots, n - q. \tag{3.2}$$

or in matrix form

$$L \cdot \mathbf{c} = 0, \tag{3.3}$$

where $L$ is a *Loewner* or *divided-differences* matrix of dimensions $(n-q) \times q$, derived from the given pairs of points. This Loewner matrix is a major instrument for the rational interpolation problem. The key property of this matrix is that its rank is directly related to the degree of the corresponding minimal-degree interpolating function. More about the Loewner matrix characteristics is found in [8].

The interpolation problem now reduces to determining the $\mathbf{c}$ vector such that (3.3) is satisfied. Once $\mathbf{c}$ is obtained, we can compute the rational interpolation function $y(x, \mathbf{c}) = \frac{b(x,\mathbf{c})}{a(x,\mathbf{c})}$ where

$$b(x, \mathbf{c}) = \sum_{i}^{n} c_i y_i \prod_{j \neq i} (x - x_j), \;\; a(x, \mathbf{c}) = \sum_{i}^{n} c_i \prod_{j \neq i} (x - x_j). \tag{3.4}$$

For the proof of (3.4) refer to [10]. After having reviewed what we need on Loewner matrices associated with the interpolation of scalar transfer functions, we turn our attention to the matrix transfer functions, and how their minimal state-space realizations are computed.

### State-Space Rational Interpolation

Let $Y(x)$ be a matrix transfer-function with a minimal state-space realization $\{A, B, C, D\}$ of the form

$$Y(x) = C^*(xI - A)^{-1}B + D. \tag{3.5}$$

Consider the array of points $P := \{(x_i, y_i), i = 1, \ldots, n\}$, with $x_i \neq x_j$, and $x_i, y_i \in \mathbb{C}$. Suppose now we partition the vector $\mathbf{x} = \{x_1, \ldots, x_n\}$ into two nonempty sets $R$ and $T$ called the *row set* and *column set* respectively. Let $R = \{r_1, r_2, \ldots, r_\gamma\}$, with $r_i \neq r_j$ for $i \neq j$, be the row set, and let $T = \{t_1, t_2, \ldots, t_\delta\}$, with $t_i \neq t_j$ for $i \neq j$, be the column set such that $R \cap T = \varnothing$.

If $Y(x)$ is given in terms of a causal transfer-function matrix with minimal state-space dimension $q$, the Loewner matrix $L$ associated with the transfer-function matrix $Y(x)$ is factored into a product of two matrices $M$ and $N$ with column and row rank q respectively,

$$\begin{aligned} L &= -\begin{bmatrix} C^*(r_1 I - A)^{-1} \\ C^*(r_2 I - A)^{-1} \\ \vdots \\ C^*(r_\gamma I - A)^{-1} \end{bmatrix} \cdot \begin{bmatrix} (t_1 I - A)^{-1}B, (t_2 I - A)^{-1}B, \\ \ldots, (t_\delta I - A)^{-1}B \end{bmatrix} \\ &= MN \end{aligned} \tag{3.6}$$

This is a state-space representation of the Loewner matrix [8]. Similar to the Hankel matrix, the Loewner matrix is divided into generalized controllability and observability matrices.

The main strategy now is to find the Loewner matrix $L$, i.e. to compute $\{A, B, C, D\}$ such that $M$ and $N$ are the generalized observability and controllability matrices. Appendix C recapitulates the major steps involved in computing the $L$ matrix as detailed in [8]. We will now formulate the rational interpolation problem to fit our aim of interpolating the HRTFs.

### 3.2.3 Experimental Setup

In our investigation, we use KEMAR HRTFs of length 512 samples measured at 44.1 KHz every 5° in the horizontal plane (0° azimuth) as well as in four other planes at elevations ±10°, ±20°. For every plane, a total of 72 HRTFs were available. The interpolation is done for one plane at a time. Figure 3.2 illustrates the interpolation procedure. After inserting all binaural responses (time-domain) into the rows of a common matrix $H$, the interpolation algorithm reads the matrix column-wise one column at a time. Every column contains a number of 72 sample values taken from every binaural impulse response. Every column in the $H$ matrix can be written as $Y(z) = c_0 + c_1 \cdot z^{-1} + c_2 \cdot z^{-2} + ... + c_n \cdot z^{-n}$ where $n = 71$. This filter possesses a minimal state-space realization $\{A, B, C, D\}$ having the form of (3.5).

We follow section 3.2.2 to evaluate the matrices $A, B, C$, and $D$ for a given column in $H$, we use them to compute the Loewner matrix according to (3.6). We

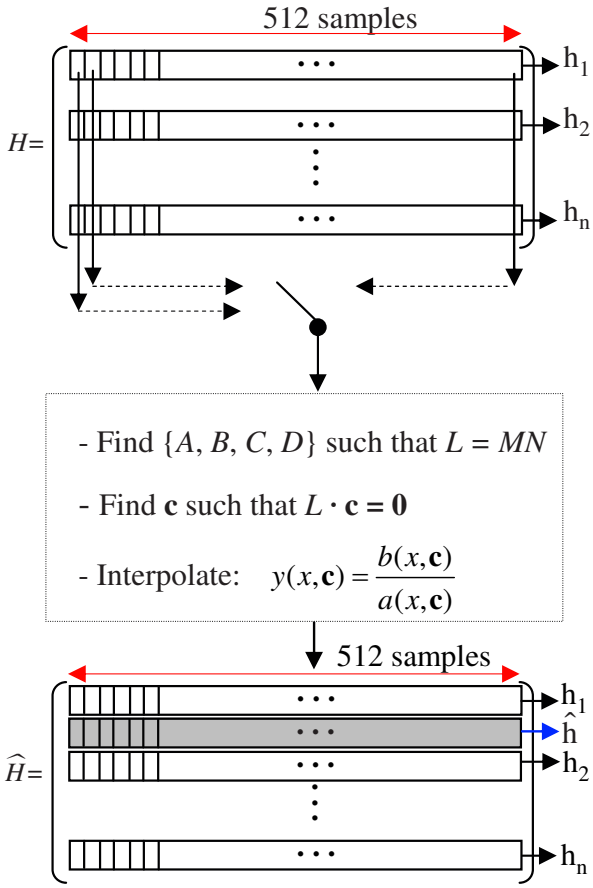


Figure 3.2: The interpolation process.

then use (3.3) to compute the vector $\mathbf{c}$ of interpolating coefficients which we insert into (3.4) to obtain the desired interpolation function $y(x, \mathbf{c})$. Using this function for non-integer values of $x$ we are able to interpolate new angles between every two adjacent points for a given column of the HRTF matrix $H$. The same process is repeated for all other columns of $H$. After running through all the 512 columns, the resulting matrix $\hat{H}$ is composed of rows that include the interpolated binaural responses.

Similar to the DFT method, in the rational interpolation method, the response to be interpolated depends on the responses at all azimuths. In the linear method, however, the binaural response to be interpolated is determined based only on two adjacent responses.

### 3.2.4 Discussion of Results

To verify the performance of the rational interpolation, we use 72 measurements for the left ear. The HRTF to be interpolated is omitted from HRTF measurements. This process is repeated for five elevation angles 0°, ±10°, ±20°. The interpolation result is compared with the corresponding available measurement and the Signal-to-Distortion Ratio (SDR) is computed,

$$SDR = 10log\frac{\sum_{n=0}^{N_h-1} h^2(n)}{\sum_{n=0}^{N_h-1}[h(n)-\hat{h}(n)]^2}$$

where $h(n)$ denotes a measured HRTF at a certain angle for the left ear, $\hat{h}(n)$ denotes an interpolated HRTF at the same angle, and $N_h$ is the HRTF length, e.g., $N_h = 512$.

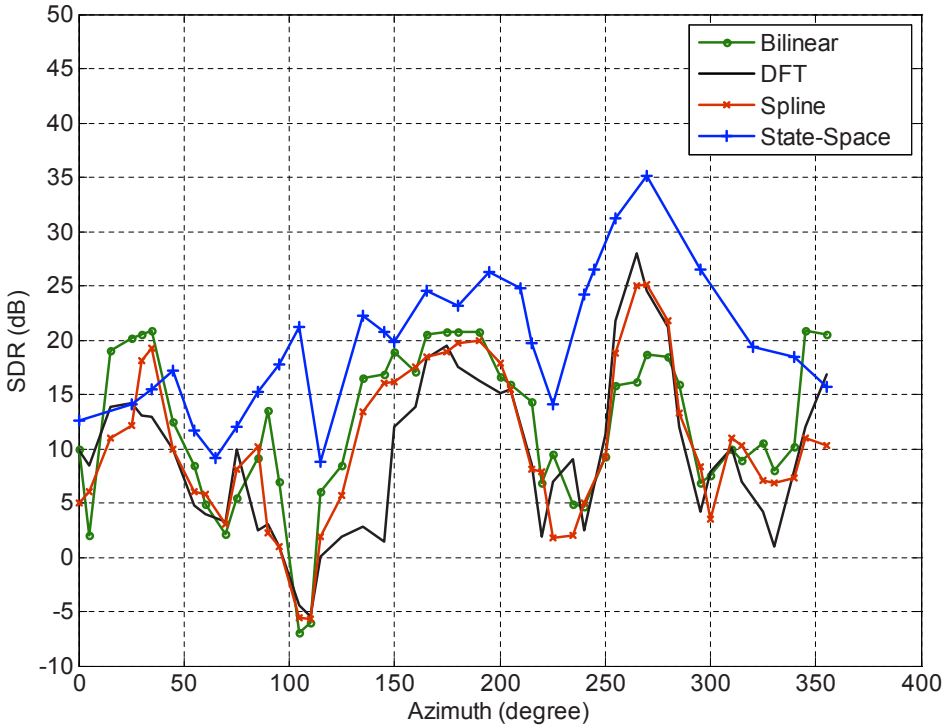


Figure 3.3: Interpolation accuracy.

Figure 3.3 shows the interpolation efficiency averaged over the five elevation planes tested. Compared with bilinear, DFT, and spline interpolation methods using time correction, the state-space rational interpolation shows comparable performance from 0° to 45° azimuths and from 345° to 360°. Over all the remaining

azimuth range the SDR for the rational method is higher than that for the other three methods. The mean SDR value for the state-space method is 28.3 dB compared to 12.07 dB, 9.46 dB and 10.44 dB for the bilinear, DFT, and spline methods, respectively.

Compared with existing interpolation techniques, the state-space method we have introduced allowed very precise reconstruction of HRTFs and proved to have higher performance for a wide range of azimuths [66], [59], [62]. Nevertheless, it should be noted that more gain can be achieved if we use the time correction method along with the presented rational interpolation method. We will use this interpolation method to increase the accuracy of a novel sound localization algorithm presented in chapter 6.

### 3.2.5 Subjective Analysis

In order to verify the theoretical results, we carried out headphone listening experiments. The goal was to study the performance of the state-of-the-art interpolation techniques as compared to our RSS method. A total of 20 male test subjects participated in the listening experiment with ages ranging between 23 and 35. The hearing of all test subjects was tested using standard audiometry. None of the subjects had reportable hearing loss that could effect the test results.

#### Test Methods

In the first test method, the subjects were asked to determine the angle of arrival of the synthesized test tones. In each trial the same test signal was repeated two times with 0.5s silence between each play. The HRTF used to synthesize the test tones was randomly chosen from 12 Kemar HRTFs measured every 30° in the horizontal plane. This type of subjective tests, however, is prone to localization errors such as in-head localization and front back confusions, therefore, a second listening test was implemented.

In this test, an A/B paired comparison hidden reference paradigm was employed. The subjects were asked to grade localization impairment against the hidden reference on a continuous 1.0 to 5.0 scale (1- very different, 2- Slightly different, 3- Slightly similar, 4- Relatively similar, 5- No difference). The hidden reference in each case was randomly chosen from the 72 measured Kemar HRTFs of the horizontal plane. The other test signal was synthesized using interpolated HRTFs. The four above-mentioned interpolation methods were tested. In each trial two test sequences were presented with 0.5s between each sequence, i.e. A/B - 0.5s - B/A. Two different random orders of presentation were used to minimize bias. Listeners were

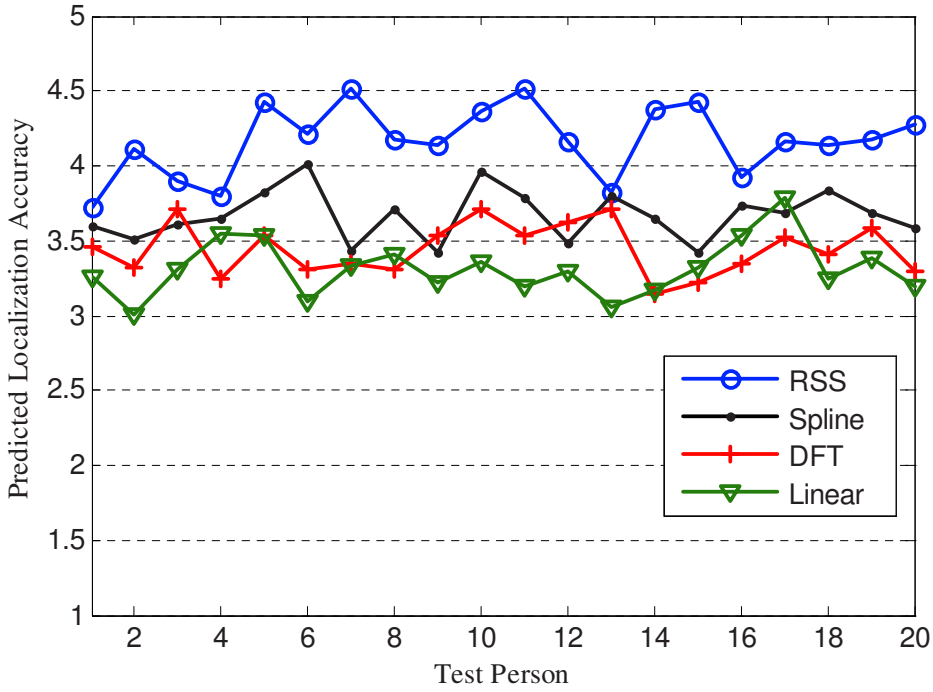


Figure 3.4: Predicted localization accuracy averaged for 20 test subjects every 30° from 0° to 330°.

given written and oral instructions

**Test Stimuli**

A pink noise[1] sample with a length of one second was used in the final experiment. The level of the stimuli was adjusted so that the peak A-weighted SPL did not exceed 70 dB at any point. This has been done in order to avoid level adaptation. No gain adjusting of the test sequences calculated for one person was carried out, since the only variability in level was introduced by the used HRTF filters.

### 3.2.6 Performance Results

The test stimuli were presented over headphones. A computer keyboard was placed in front of the test person. Each test person was individually familiarized and

[1]Pink noise or 1/f noise is a signal whose power spectral density is proportional to the reciprocal of the frequency. The name arises from being intermediate between white noise (1/f0) and red noise (1/f2), more commonly known as Brownian noise.

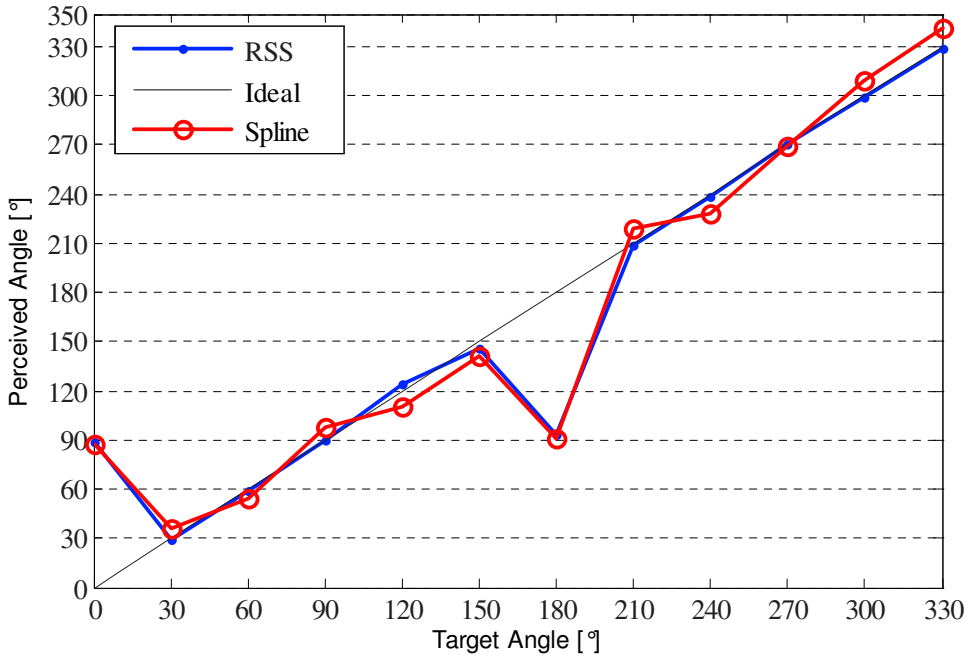


Figure 3.5: Listening test results averaged over 20 subjects: perceived angle versus target angle.

instructed to grade the localization scales for each test signal pair. The results of the listening tests were gathered automatically by a program and transferred into a statistical software package where analysis was performed.

For the first test, every subject was asked to predict an angle of arrival for each of the 12 test tones synthesized using HRTFs interpolated every 30° in the horizontal plane. The HRTFs were interpolated using the above mentioned interpolation techniques.For every synthesized test tone, i.e. for every target angle, the predicted values were averaged over 20 subjects. Figure 3.4 summarizes the average predicted angle of arrival vs. the target angle. The HRTFs used in tone synthesis where interpolated using the RSS and Spline; the linear and DFT methods performed similar to Spline and were therefore not included in Figure 3.4 for clarity. One can observe that the average localization accuracy has a wider confidence interval for HRTFs obtained using Spline interpolation as compared to the RSS method. Furthermore, for the interval 240° to 310°, the subjects reported highest localization accuracy when the RSS method was deployed. This result is conforming with Figure 3.2 where for the same angular interval, the interpolated HRTF has SDR values above 40dB. In addition, the front back confusion problem is clearly noticeable at both 0° and 180°, i.e. at those locations where the sound source is directly in front or behind the listener.

Table 3.2: Performance comparison in terms of million instructions per second (MIPS).

| | Interpolation | Convolution | MIPS |
|---|---|---|---|
| **RSS** | 4,722,135 | 5,564,146 | 10.28 |
| **Linear** | 906,217 | 5,564,146 | 6.47 |
| **DFT** | 2,116,354 | 5,564,146 | 7.68 |
| **Spline** | 1,348,091 | 5,564,146 | 6.91 |

In the second test, the subjects were not asked to determine the angle of arrival of the test tones. This rules out the possibility of checking for localization errors such as in-head localization or front back confusions, especially that we are using Kemar HRTFs, i.e. non individualized HRTFs. Therefore we made the assumption that the reference HRTF is "perfect" and the quality degradation is always related to that reference. Figure 3.5 summarizes the average localization accuracy for every test person. Complying with the theoretical results, the RSS interpolation method yields better localization results than the state-of-the-art methods and consequently insures high-fidelity reproduction of the HRTFs.

Finally, for the verification of real-time implementation, a comparison of methods in terms of CPU usage, e.g. million instructions per second (MIPS), is shown in Table 6.1. The table displays the number of instructions used for processing 1 second of input data. The kernel of suggested RSS algorithm requires 10.28 MIPS for the input data sampled at 44.1kHz. While this value exceeds the processing power of the other interpolation methods, it remains well below the real-time processing bound of our 32-bit Pentium IV, 1.4 GHz platform s [69].

## 3.3 Efficient State-Space HRTF Inversion

Our approach for sound localization depends on finding the inverse filters of the HRTFs, and saving them for later use. The inverse filter was made available by simply exchanging the values of the numerator and the denominator, i.e., for the FIR filter $F(z) = c_0 + c_1 \cdot z^{-1} + c_2 \cdot z^{-1} + \ldots + c_n z^{-n}$, the inverse filter would be $G(z) = \frac{1}{F(z)}$. Similarly, this concept applies for the IIR filter.

However, direct methods to invert FIR filter transfer functions may produce unstable filters. This is the case with the inverted HRTF filters, using the FFT method, especially because all HRTFs include a *linear-phase component*, i.e. pure delay, which is vital for maintaining the correct inter-aural time difference. The task is then to take the non minimum-phase filter and determine a corresponding

minimum-phase filter, which can then be safely inverted to produce a stable inverse filter. The resulting filter may be non-causal but serves as an approximation to the original unstable filter. Non-causality is a small price to pay if magnitude and phase information are critical to performance, which is the case in our situation. The following section will provide details of the inversion process we have adopted in this work using an inner-outer factorization approach.

One method which handles the previously mentioned inversion problem ensuring stability is to replace the unstable inverse by an anti-causal yet bounded inverse. This can be done efficiently using the state-space inversion method with inner-outer factorization. According to this method, the inner factor captures the part of the transfer function that causes the instability in the inverse, while the outer part can be straightforwardly inverted.

### 3.3.1 Problem Formulation

There are a variety of reasons why a state-space representation of the HRTFs is beneficial. While transfer functions only deal with input/output behavior or the system, state-space forms provide an easy access to the internal features and response of the system. General system properties, for example, the system controllability or observability can be defined and determined. One of the main advantages of the state-space modeling is the possibility of ordering the system states based on their significance to the representation of the system characteristics. This is a very important feature that allows model simplification, e.g. HRTF order reduction, by direct truncation, i.e., by discarding its "less important" states.

For rational time-invariant single-input single-output systems, the inner-outer factorization is a factorization of an analytical (causal) transfer function $H(z)$ into the product of an inner and an outer transfer function according to

$$H(z) = H_o(z)V(z). \tag{3.7}$$

The inner factor $V(z)$ has its poles outside the unit disc and has modulus 1 on the unit circle, whereas the outer factor $H_o(z)$ and its inverse are analytical in the open unit disc. Before deriving these inner and outer factors, we will review the steps involved in inverting a given transfer function given a state-space representation thereof.

Let $H(z) = D + Cz(I - Az)^{-1}B$ be a minimal state-space representation of a given HRTF, with $D$ being square and non-singular. This transfer function can also be represented by the state-space equations,

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k + Du_k \end{aligned} \tag{3.8}$$

where $x$ is the $n$-dimensional state vector, $u$ is the scalar system input, $y$ its output, with the matrix A, vectors B and C, and scalar D fully defining the system.

Starting with the transfer function of the general state-space model, $H(z) = D + Cz(I - Az)^{-1}B$, we may first observe that the poles of $H(z)$ are either the same as or some subset of the poles of $H_p(z) = z(I - Az)^{-1}$ (They are the same when all modes are controllable and observable). By Cramer's rule for matrix inversion, the denominator polynomial for $H_p(z)$ is given by the determinant $D(z) = |(Iz^{-1} - A)^{-1}|$, where $|.|$ denotes the determinant of a square matrix. In linear algebra, the polynomial $D(z) = |(Iz^{-1} - A)^{-1}|$ is called the characteristic polynomial for the matrix $A$. The roots of the characteristic polynomial are the eigenvalues of $A$. Thus, the eigenvalues of the state transition matrix are the poles of the corresponding linear time-invariant system. In particular, note that the poles of the system do not depend on the matrices $B$, $C$, and $D$, although these matrices, by placing system zeros, can cause pole-zero cancelations (unobservable or uncontrollable modes).

The direct way of inverting the transfer function $H(z)$ results in an unstable system, knowing that $H(z)$ represents a non minimum-phase transfer function. This unstable inverse could be, in state-space, directly derived. We first take the second part of (3.8) and solve for $u_k$:

$$u_k = D^{-1}y_k - D^{-1}Cx_k. \tag{3.9}$$

Inserting this in the first part of (3.8)

$$x_{k+1} = Ax_k + B(D^{-1}y_k - D^{-1}Cx_k), \tag{3.10}$$

leads to the inversion in the state-space, which can be written as

$$\begin{aligned} x_{k+1} = &\quad (A - BD^{-1}C)x_k &+ &\quad BD^{-1}y_k \\ u_k = &\quad -D^{-1}Cx_k &+ &\quad D^{-1}y_k \end{aligned} \tag{3.11}$$

We denote the quadruple $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$, where $\tilde{A} = A - BD^{-1}C$, $\tilde{B} = BD^{-1}$, $\tilde{C} = -D^{-1}C$, and $\tilde{D} = D^{-1}$, as the state-space realization of the unstable inverse transfer function, $H^{-1}(z)$, corresponding to the inverse HRTF. The state-space realization $\tilde{A}$ has its poles outside the unit circle and, therefore, drives the system unstable. To ensure stability, we implement the inner-outer factorization theorem stated below.

### 3.3.2 Inner-Outer Factorization

Given a minimal state-space realization of the transfer function $H(z)$, we would like to find the factors $V(z)$ and $H_o(z)$, as in (3.7), where $V(z)$ is unitary and $H_o(z)$ is

an outer function, that is to say it is minimum phase, and hence $H_o^{-1}(z)$ is bounded but not causal.

Equation (3.7) can be expressed in a state-space form:

$$D + Cz(I - Az)^{-1}B = \tag{3.12}$$

$$\left[D_o + C_o z(I - Az)^{-1}B\right]\left[D_v + C_v z(I - A_v z)^{-1}B_v\right] \tag{3.13}$$

where $\{A_v, B_v, C_v, D_v\}$ is a realization for $V(z)$, and $\{A, B, C_0, D_0\}$ is a realization for $H_o(z)$.

Expansion of the quadratic term in (3.12), and equating members leads to

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} Y & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} Y & C_0 \\ 0 & D_0 \end{bmatrix}\begin{bmatrix} A_v & B_v \\ C_v & D_v \end{bmatrix} \tag{3.14}$$

where the diagonal matrix $Y$ satisfies the Lyapunov-Stein equation

$$Y = BB_v^\star + AYA_v^\star \tag{3.15}$$

To get the inner and outer factors we are looking for, Eq. (3.15) must be solvable, and the kernel and maximality requirements on $Y$ and $D_o$ must indeed produce an outer factor $H_o(z)$.

*Theorem*: Let $W$ be a unitary matrix, and $Y$ be a uniformly bounded matrix which satisfies the following equality,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} Y & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} Y & B_0 \\ 0 & D_0 \end{bmatrix} W \tag{3.16}$$

such that $Y$ has a maximal dimension and $\ker(Y.) = 0$. Let

$$W = \begin{bmatrix} A_v & B_v \\ C_v & D_v \end{bmatrix}. \tag{3.17}$$

Then $\{A_v, B_v, C_v, D_v\}$ is an isometric realization for the sought inner factor $V(z)$, and $\{A, B, C_0, D_0\}$ is a realization for the outer factor $H_o(z)$. The proof of the above theorem is detailed in [33]. In appendix A, we recapitulate the main steps of this theorem.

We shall use the above-mentioned state-space inversion algorithm to considerably stabilize and enhance the performance of our new sound localization techniques presented in the following chapter.

# Chapter 4

# Enhanced Sound Source Localization Techniques

In this chapter, we have considerably improved the initial matched filtering approach to sound localization in order to achieve low-complexity tracking of moving sound. the humanoid detects the current location of the sound source using a number of correlation operations between the input signals at the microphones and a generic set of transfer functions. Using a properly tuned Kalman filter, a Region Of Interest (ROI) is automatically extracted within the HRTFs, leading to a faster detection, due to less correlation operations and low processing power requirements. The proposed method is demonstrated through simulations and is further tested in a household environment. In contrast to microphone-array methods, using only two microphones, the new system demonstrated high precision 3D sound tracking and enabled a low-complexity implementation on the humanoid DSP platform [66].

## 4.1 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 [60]. However, a more direct approach can be taken to localize a sound source. Instead of attempting to 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

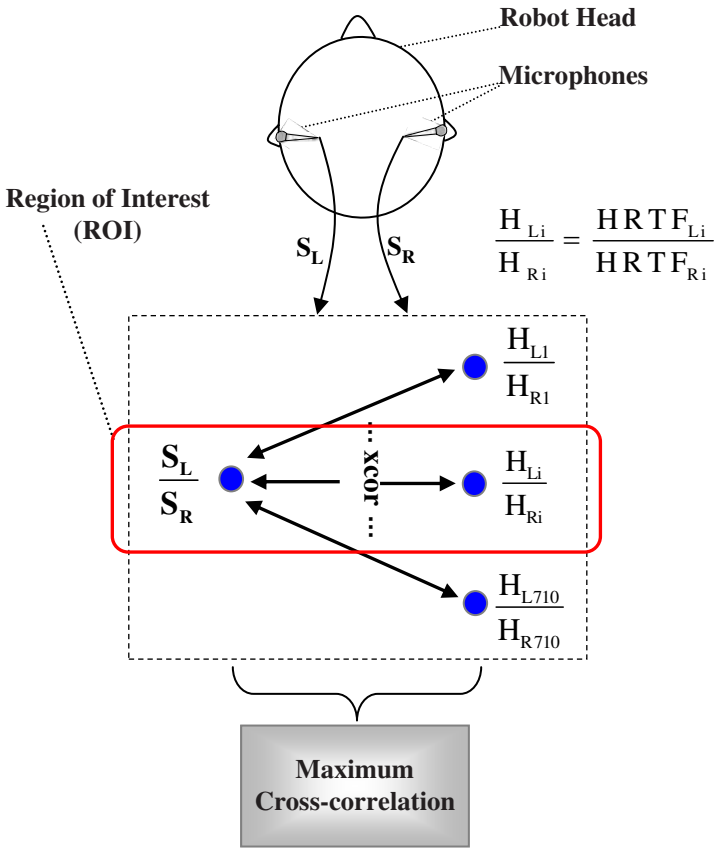


Figure 4.1: Flow chart of the Source Cancelation Algorithm (SCA).

4.1.

Basically, the received signals at the microphones 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. Like this, the SCA depends only on the correlation factor between incoming and saved HRTFs ratios. Hence, the SCA outperforms the previously proposed method as it is independent from the characteristics of the impinging sound sources on the artificial ears and torso, which ensures more stability and more tolerability to noise and reverberations.

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 SCA 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.

### 4.1.1 Kalman Filtering and ROI Extraction

Although applying appropriate reduction techniques, as in chapter 2, the length of the impulse responses can be reduced to a hundred or even fewer samples, thus reducing the overall localization time, the HRTF database could be very dense, and the convolution with all possible HRTFs in the database becomes computationally exhaustive. To solve this problem, especially for moving sound sources, a Kalman filter is tailored to predict a ROI, the sound source might be heading to, according to some movement models. Therefore, a quick search for the correct HRTF pair within a small ROI is now ensured, and, consequently, a very fast tracking of the moving sound trajectory [74]. The workflow of the SCA attached to a Kalman filter is depicted in Figure 4.2.

The sound localization algorithm initializes by making a search in the whole HRTF dataset looking for the starting position of the sound source. Once the initial position is pinpointed, an initial ROI is localized around this initial position. Then, the source starts moving, and a new ROI is identified and automatically updated into the system. The Kalman filter used for the ROI updating consists of a set of mathematical recursive algorithms computationally capable of predicting the future state of a process by minimizing a mean of the squared error between the measurements and the predictions of the moving sound locations. A detailed description can be found in [132].

The Kalman filter we are utilizing [65], applies to a linear dynamical system, the state space model of which consists of two equations:

$$x_{k+1} = Ax_k + Bu_k + w \tag{4.1}$$

$$y_{k+1} = Cx_{k+1} + v \tag{4.2}$$

Equations 4.1 and 4.2 are called the process and the measurement equations respectively. The variable $x \in \Re^n$ is the state of the discrete-time system and $y \in \Re^m$ is the system's output such as positions or angles depending on the movement model. The variable $u$ models the sound source velocity. The random variable $w$ and $v$ represent the white gaussian process and measurement noise, respectively. The $n \times n$ matrix $A$ relates the state at the current time step $k$ to the state at the future step

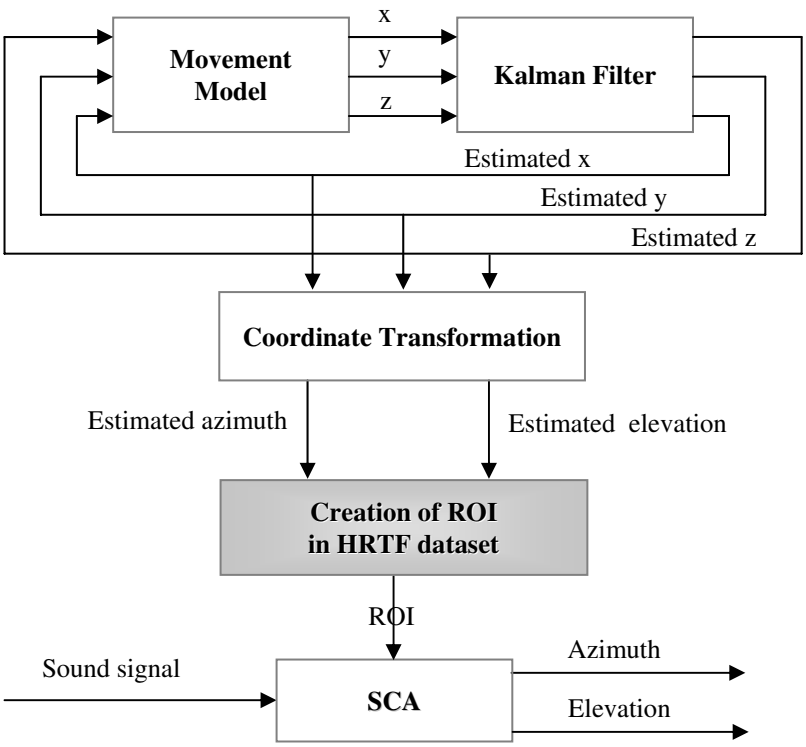


Figure 4.2: Flowchart of the Source Cancelation Algorithm using a Region of Interest (ROI).

$k+1$. The $n \times l$ matrix $B$ relates the optional control input $u \in \Re^l$ to the state $x$. The $m \times n$ matrix $C$ relates the state to the measurement $y_{k+1}$.

The linear model we have adopted corresponds to a sound source moving with constant velocity. This velocity is incorporated within the state vector, $x(k) = (x, y, z, \dot{x}, \dot{y}, \dot{z})^T$, by taking the derivative with respect to time, $\dot{x} = \frac{dx}{dt}$, $\dot{y} = \frac{dy}{dt}$ and $\dot{z} = \frac{dz}{dt}$. According to the movement equations, based on Newton, the sampling time $T$ for each coordinate $\Lambda \in \{x, y, z\}$ is calculated for the transition from $T_k$ to $T_{k+1}$:

$$\Lambda_{k+1} = \Lambda_k + T\dot{\Lambda}_k \tag{4.3}$$

$$\dot{\Lambda}_{k+1} = \dot{\Lambda}_k \tag{4.4}$$

The cartesian coordinates, provided by the Kalman filter using the above-mentioned model, are transformed to spherical coordinates to stay compatible with the SCA azimuth and elevation coordinates.

The above mentioned localization techniques are simulated and further tested in a real-life environment. A KEMAR head mounted on an artificial torso and equipped with two small microhpones and silicon outer ears is available for the tests.

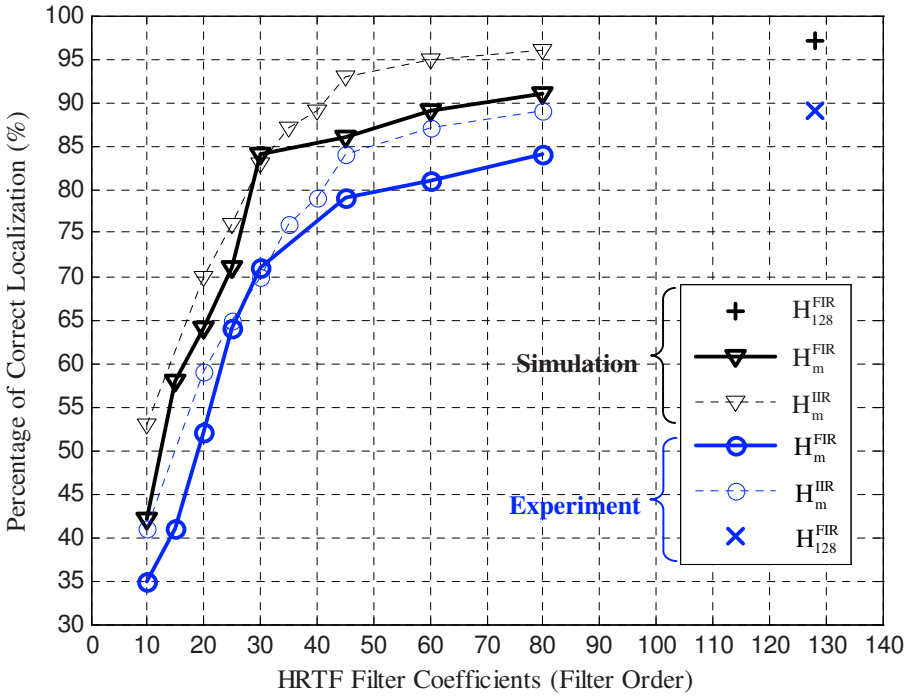


Figure 4.3: Percentage of correct localization using DFE, PCA and BMT reduced HRTFs.

Beside the normal testing procedure, a case study is presented for the situation where the humanoid is to detect sound without using its pinnae, i.e. the silicon outer ears are taken away.

### 4.1.2 Simulation 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 locations in the 3D space. 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 710 reduced and state-space inverted HRTFs. See chapter 3, section 3.3 for more details about the state-space HRTF inversion. 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 correlation factor. The reduction techniques, namely DFE,

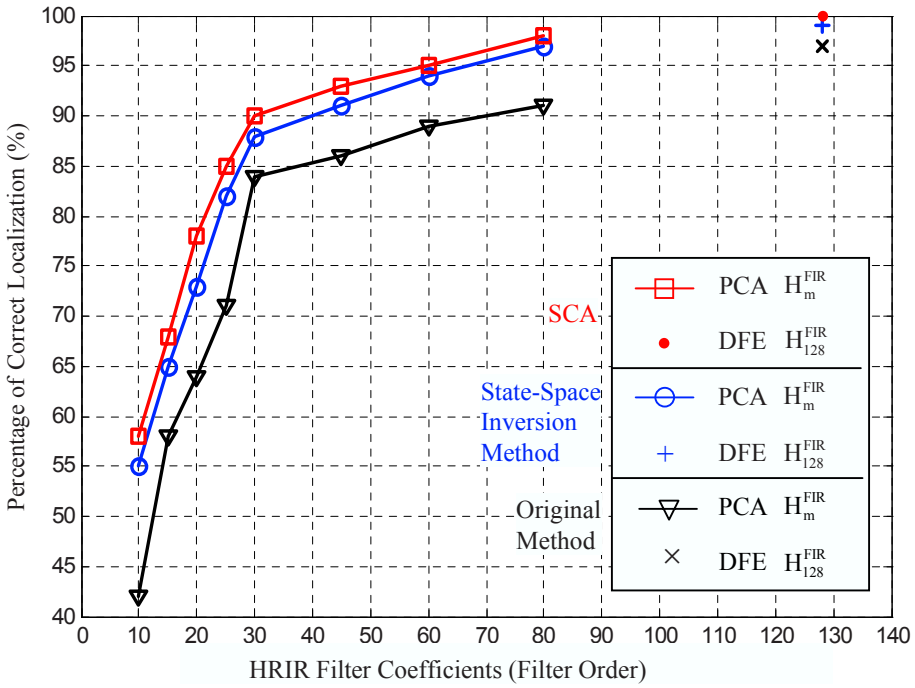


Figure 4.4: Percentage of correct localization using SCA compared to the state-space inversion, and to the original method in [66].

PCA and BMT were used to create two different reduced models of the original HRTFs. The performance of each of these models is illustrated in Figure 4.3.

To simulate the reverberation in our room environment, the image method for room acoustics was used [7]. A reverberation time of $RT = 0.21s$ was chosen. The simulation setup and room dimensions were defined to match the experimental room environment. 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 source's and microphone's positions. After recombining the convolution results, random Gaussian noise was finally added to each microphone signal yielding an SNR level of 20dB.

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 filter coefficients. Most significantly, it was observed that, up to order 30 PCA-reduced and order 35 BMT-reduced HRTFs, all the falsely localized angles fall within the very close vicinity of the original sound locations.

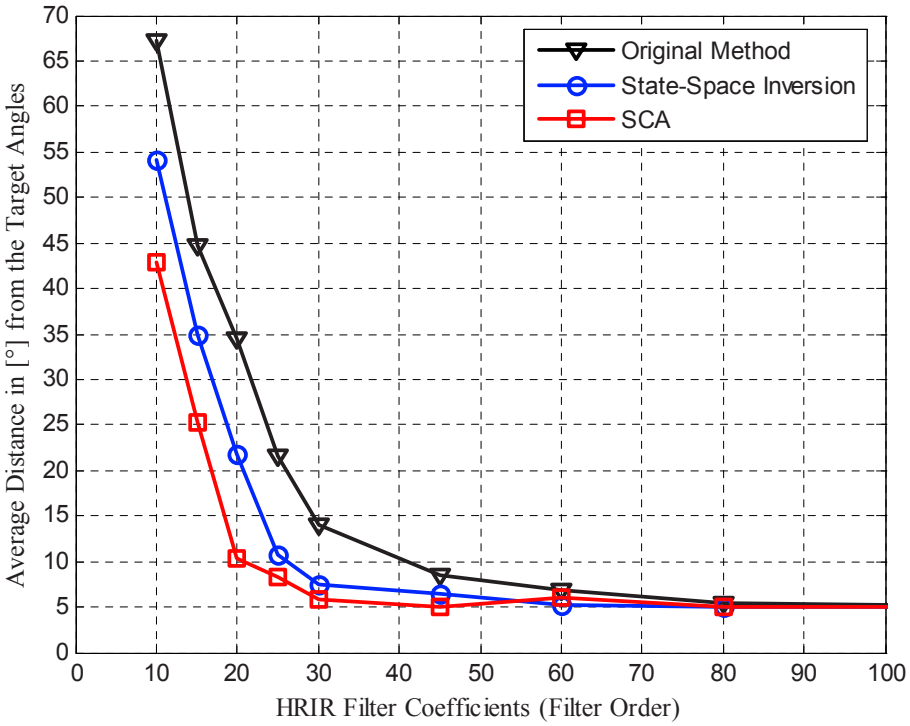


Figure 4.5: The falsely localized sound sources: average distance to their target positions, for every HRIR length.

The performance of the original method and the enhanced one, i.e. SCA algorithm, with state-space inversion is illustrated in Figure 4.4. The solid line with square markers, as well as the dot at the top right side of the figure, show the SCA percentage of correct localization versus the length of the HRTF in samples. The solid line with circle markers, as well as the plus sign, show the state-space inversion percentage of correct localization. The solid line with triangle markers, and the multiplication sign, refer to the previous FFT method performance [82, 66].

Using the diffuse-field equalized 128-sample HRTFs, $\mathrm{H}_{128}^{FIR}$, the SCA percentage of correct localization is 100 %, this means that all the 100 locations where perfectly detected at their target 3D location. The state-space inversion algorithm correctly locate 99% of the sources, compared to 96% for the previous method using FFT inversion.

Using the PCA-reduced set, $\mathrm{H}_{m}^{FIR}$, the SCA percentage of correct localization falls between 58% to 98%. The state-space inversion localization percentage is lower and falls between 55% to 97% compared to 42% and 91% for the previous FFT method, with the HRTF being within 10 to 45 samples, i.e. $10 \leq m \leq 45$. It should be noted that, while using 35 PCA-reduced HRTFs, all of the falsely localized

angles 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 Figure 4.5. Intuitively, with more HRIR samples, the average distance to the target sound source location decreases. Note that the minimum distance to the target position is 5°, and this is due to the fact that the minimum angle between the HRTFs of the database we are using is 5°. Obviously, if we use a densely sampled database, e.g. with a HRTF every 1°, the average distance to the target sound locations is expected to notably decrease.

### 4.1.3 Experimental Results

In our household experimental setup, 100 binaural different recordings of a broadband sound signal, placed at different angles around a artificial head, were obtained using a artificial head and torso with two artificial ears in a reverberant room. Our hardware setup is illustrated in Figure 4.6.

The speaker was held at a constant distance of 1.3 meters from the head. The recording environment was a laboratory room measuring where the walls, ceiling, and Floor are made of unpainted concrete. One wall has a $5m \times 2m$ glass window and is facing the dummy head. The dummy head and torso are placed on a rotating table in the middle of the room. The dummy head artificial ears and microphones are held at a constant height of 1.5 meters from the floor. The room contains objects like tables, chairs, and computer screens.

The level of reverberation in the room was experimentally measured by means of a loudspeaker emitting a high level white noise signal. Measuring the 60dB decay period of the sound pressure level after the source signal is switched off, for a number of speaker and microphone positions, provided the frequency-averaged reverberation time $RT = 0.21s$.

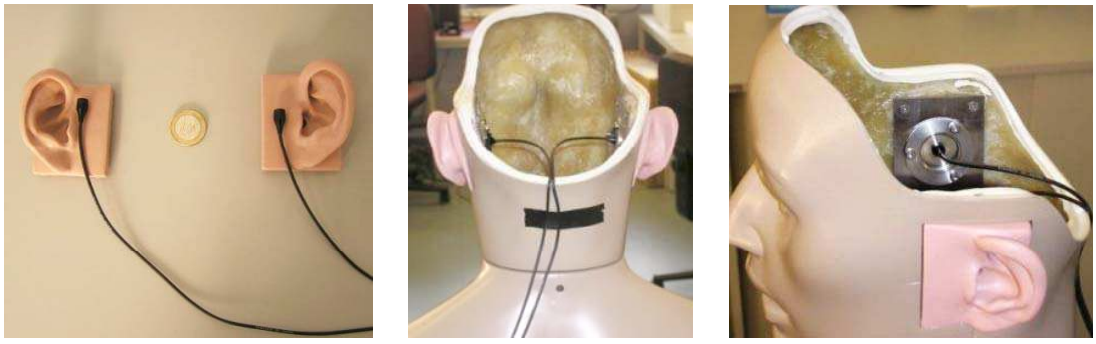

Figure 4.6: The laboratory hardware setup.

To keep a fair comparison with the simulation setup, each of the recordings was 350 msecs long. 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 the SCA, the state-space inversion algorithm, and the original algorithm. A HRTF database reduced using the DFE method, $H_{128}^{FIR}$, was available for the test.

Using the original method, 80% of the estimated angles are correctly localized, compared to 97% for the simulated case. This difference is mainly due to the external reverberation, and internal equipment noise, and due to the differences between the artificial manikin model used in the experiment and the KEMAR model used to obtain the HRTF dataset. The 20% falsely localized sound sources are located at an average angular distance of 27.5° from the target angle.

Using state-space inversion, 88% of the estimated angles are found at the correct location, compared with the 99% obtained from the theoretical simulation. The 14% falsely localized sources are found at an average angular distance of 11.45° of the target sound source location. This is a remarkable improvement, compared to the repartition of the falsely localized angles in the case of the original method using the FFT inversion.

Using the SCA, 94% of the estimated azimuth and elevation angles turned out to be exactly at the target location, compared to 100% obtained from the simulation. The other falsely localized 6% were identified at the near vicinity of the target angles, with an average distance of 5.85°. Again, this is a notable observation, and a considerable improvement compared to the original as well as the state-space inversion algorithms. As foreseen in our theoretical study, the SCA outperforms, in simulations as well as in real life, the previously proposed methods. This is due to the fact that SCA is canceling the sound source signals, and thus the accompanying noise, making it, thus, less dependent on the characteristics of the sound sources, and consequently more stable and more tolerable to noise and reverberations.

### 4.1.4 A Case Study

Finally, a case study is performed to test the SCA algorithm in the special case where the two pinnae of the artificial head are taken away. The microphones are placed at a distance of 26 mm away from the ear canal opening. The performance in this case is depicted in Figure 5.8. For comparison purposes, the SCA theoretical and experimental performance for the artificial head deployed with two pinnae are also shown. The line with circle markers represents the theoretical performance of the SCA using PCA-reduced HRTFs, and the plus sign represents the performance using 128-sample long DFE-reduced HRTFs. The line with circle markers, and

the plus sign, correspond to the experimental setup, where the artificial head is equipped with two artificial pinnae. The line at the bottom of the figure, and the multiplication sign, represent the performance of the SCA in the no-pinnae case.

The degraded performance of the SCA in the no-pinnae case is obvious. The pinnae consist of asymmetrical grooves and notches which accentuate or suppress the mid and high frequency energy content of the sound spectrum to a certain degree, depending very much on both the location and frequency content of the sound source. These filtering effects are embedded within the HRTFs. Essentially, the HRTF modifies the spectrum and timing of a sound signal reaching the ears in a location dependent manner which is recognized by the listener and used as a localization cue. When the pinnae are taken away, the incoming signal lacks the HRTF filtering effects and the SCA fails.

If the HRTFs were initially measured using a artificial head without pinnae, the localization accuracy would be expected to increase, but only to some extent, since without pinnae, many problems will emerge, especially the famous front-back confusion ambiguity. As early as 1967, the filtering effects introduced by the pinnae were given great importance, especially in the work of Bateau [12]. This work

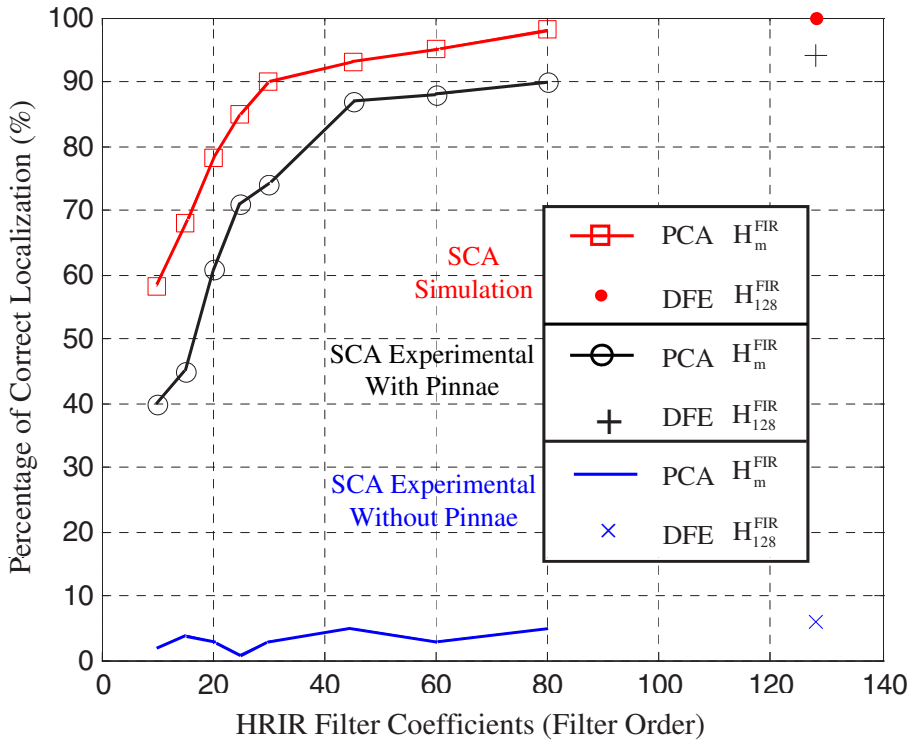


Figure 4.7: Case study: The SCA theoretical and experimental performance using a artificial head with or without pinnae.

demonstrated that the duplex theory, which was prevailing at that time, and which explains the human hearing localization based solely on the two ITD and ILD cues, is insufficient in explaining the human spatial hearing, and in resolving ambiguous situations such as the front-back confusion problem.

### 4.1.5 Performance Comparison

We evaluate the performance of our SCA algorithm in comparison to the well-known generalized correlation method described in chapter 1. The method applies Fourier transform to both microphone signals and uses weighting functions to accentuate the signal passed to a correlator for those frequencies at which the coherence or signal-to-noise ration(SNR) is highest. In our experimental setup, 100 binaural different recordings of a broadband sound signal, placed at different angles around a artificial head, were obtained using KEMAR. Table 4.1 shows different correlation-based localization methods and their average deviation to the target sound source location.

The table shows that the localization error for the direct correlation method without a weighting function is the highest. This is obvious in our reverberant environment. The localization error decreases when applying SCOT, PHAT, or ML and reaches a minimum with the SCA algorithm. It should be noted that the above mentioned correlation methods using only two microphones fail to detect sound in three dimensions and suffer from the front/back confusion problem. Conversely, using two microphones, the SCA method is not restricted to azimuthal localization, it covers all the three dimensional space.

Moreover, it should be noted that 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.

Hence, we have applied the new HRTF interpolation method [66] proposed in chapter 3 to obtain a high-spatial-resolution HRTF database with one HRTF every 1° azimuth, spanning an elevation range from -20° to 60°. Using this database, we compare our method with the 3D robotic sound localization system proposed in [48].

This method uses a spherical robot head with spiral-formed ears of slightly different sizes. The spiral form makes it easy to mathematically derive the IID, ITD and spectral cues used for localization. The robot was made to learn these cues by supervised learning or by using vision. A total of 132 sound recordings were taken

Table 4.1: Performance comparison with generalized cross correlation methods.

| **Window** | **Scope** | **Mean Localization Error** |
|---|---|---|
| **KORR** | direct correlation without a window | 15.8° |
| **SCOT (Smoothed Coherence Transform)** | suppresses tonal fractions | 11.4° |
| **PHAT (PHAse Transform)** | uses only the phase of the cross spectrum | 9.0° |
| **ML (Maximum-Likelihood)** | minimizes the variance of the time delay estimation | 8.5° |
| **SCA (Source Cancelation Algorithm)** | HRTF-based localization | 5° |

in a silent room with a white-noise sound source located 1.5 meters from the robot. Each of the recordings had a duration of 1 second. These recordings were used to extract a database of features consisting of ITDs and notch information. This database was later used in a localization experiment done in a real environment.

In this case a speech signal is played at different angles and at a distance of 1.5 meters from the robot head. Within the range of head movements happening in the experiment, i.e. from -30 to 30 degrees, the average error in the estimated azimuths and elevations is 5.7 degrees. Outside this range, however, the method undergoes front-back ambiguities and the localization errors increase considerably. Compared to this method, our SCA localization algorithm exhibits an average angular error of 2.5° for speech signals and does not require any supervised learning or vision.

### 4.1.6 Region of Interest

In our second experimental setup, several binaural recordings of a broadband sound source were taken. In every recording, the source is moving at 10°/sec, in the zero-elevation plain, and is following a circular trajectory 2 meters away from the humanoid head located in the center. The recorded sound signals, also containing external and electronic noise, were used as inputs to our sound localization algorithm. The sound localization algorithm depicted in Figure 4.1 initializes by searching the 710 HRTFs, looking for the starting position of the sound source. Once the initial position is pinpointed, an initial ROI is localized around this position. Next, the source starts moving, and a new ROI is identified and automatically updated using

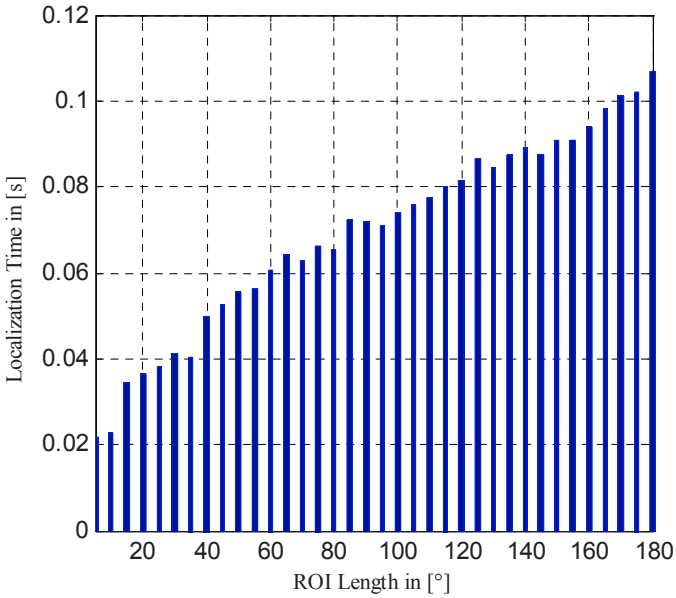


Figure 4.8: SCA processing time for different ROI intervals.

a Kalman filter. This way, the sound localization operation, particularly the correlation process, proceeds only in this ROI instead of searching the whole HRTF dataset. Thus, a considerable and dispensable computation time is avoided. The size of the generated ROI depends on the sound source's velocity and acceleration and on the model's sampling time. Obviously, the faster the source moves, and the greater the sampling time is, the bigger the ROI area becomes. To evaluate the performance of the new sound tracking setup, we set the SCA algorithm to work with BMT-reduced HRTFs of order 45. We measured the total time taken by the tracking algorithm to localize the source at one instant in time, for several ROI lengths.Without a ROI, the SCA runs through all 710 HRTFs and requires an average processing time of 0.741 sec to detect the location of a sound at a certain location. This localization time dropped down to 0.052 sec on average by using a ROI 50° long in azimuth. Further processing-time results, corresponding to diverse ROI lengths, are depicted in Figure 4.8. The algorithm was implemented in Matlab and runs on a pentium IV 1.4 Ghz processor. Using a simple Kalman filter for predicting appropriate ROIs, we attained a processing-time reduction of more than 90% as a result of less convolution and correlation operations. These computation reductions ensure a very quick tracking behavior [75].

## 4.2 Cross Convolution Approach

The matched filtering approach and SCA algorithm require the availability of the inverse HRTFs. The problem arises when the inverted filter is an unstable one. This is the case with the inverted HRTF filters especially that all HRTFs include a linear-phase component, i.e. pure delay, which is vital for maintaining the correct inter-aural time difference. Furthermore, both SCA and matched filtering assume a non-reverberant environment. Hence, we propose an algorithm which is robust to environmental noise and reverberations and which do not involve calculation of inverse HRTFs [85].

The flowchart for the convolution based HRTF algorithm is shown in Fig. 4.9. The $\otimes$ symbol in the figure indicates the convolution operator. The total number of HRTF pairs are assumed to be $N$. The left and right received signals could be modeled as the original signal convolved with the left and right HRTFs corresponding to the direction of the sound source. If we convolve the left and right received signals with the right and left HRTFs corresponding to the source direction, the convolution results should closely match. Mathematically,this can be expressed as:

$$\mathbf{S}_L \otimes \mathbf{H}_R = (\mathbf{S} \otimes \mathbf{H}_L) \otimes \mathbf{H}_R \tag{4.5}$$

$$\mathbf{S}_R \otimes \mathbf{H}_L = (\mathbf{S} \otimes \mathbf{H}_R) \otimes \mathbf{H}_L \tag{4.6}$$

where $\otimes$ indicates the convolution operator. The term $\mathbf{S}$ is the original signal coming from a specific direction, and $\mathbf{S}_L$ and $\mathbf{S}_R$ are the signals received at left and right microphones respectively. The two terms $\mathbf{H}_L$ and $\mathbf{H}_R$ are the left and right HRTFs respectively. Fast Fourier transform is used to reduce the computational complexity of convolution in time domain.

The convolution based algorithm operates as follows: In order to determine the direction of arrival of sound, the left and right microphone signals, $\mathbf{S}_L$ and $\mathbf{S}_R$ must be filtered by all $N$ right and left HRTFs. The HRTFs that result in a pair of signals that closely resemble each other should correspond to the direction of the sound source. The direction of the sound source is assumed to be the HRTF pair with the highest correlation [86].

In order to compare the performance of SCA and convolution based HRTF algorithm in a reverberant simulated environments, a total of 100 KEMAR HRTFs corresponding to different locations in 3D space were randomly chosen and test signals were synthesized by convolving a broadband speech signal with these HRTFs. For each test signal, we processed 512 samples of the speech signal and we performed two types of tests. In the first case, we considered the reflections of the source signal to be the major factor affecting the binaural signals. For this purpose, to simulate

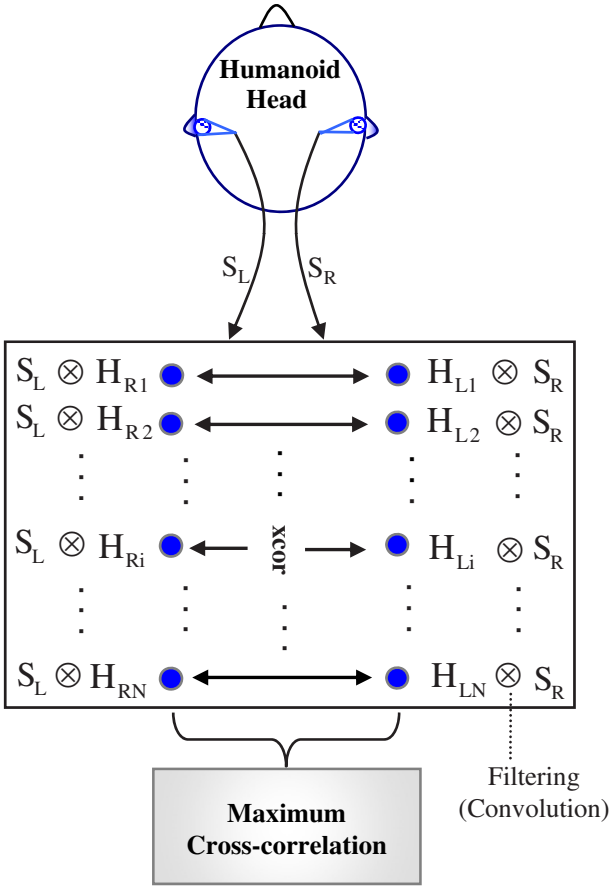


Figure 4.9: Flow chart of the convolution based algorithm.

the reverberation in our room environment, the image method for room acoustics was used [7]. In the second case, broadband noise was the major affecting factor. The top subplot in Fig. 4.10 shows the percentage of correct localization versus the reflection to signal ratio for both convolution based and SCA methods. The bottom subplot depicts the noise to signal ratio behavior. Both methods share a similar performance for reflection to signal ratio less than 0.3 and for a noise to signal power ratio less than 0.15. However, for higher noise and reverberation levels, the convolution based method outperforms the SCA algorithm. This is due to the fact that the inverse filtering operation in SCA causes the inverses to explode since the HRTFs are not minimum-phase filters.

The cross convolution algorithm, however, have to run through the correlation process for $N = 710$ HRTF pairs every time a sound source is to be localized. This increases the computational complexity and prohibits real time performance. In real environments, there are several effects such as noise and reverberations due to

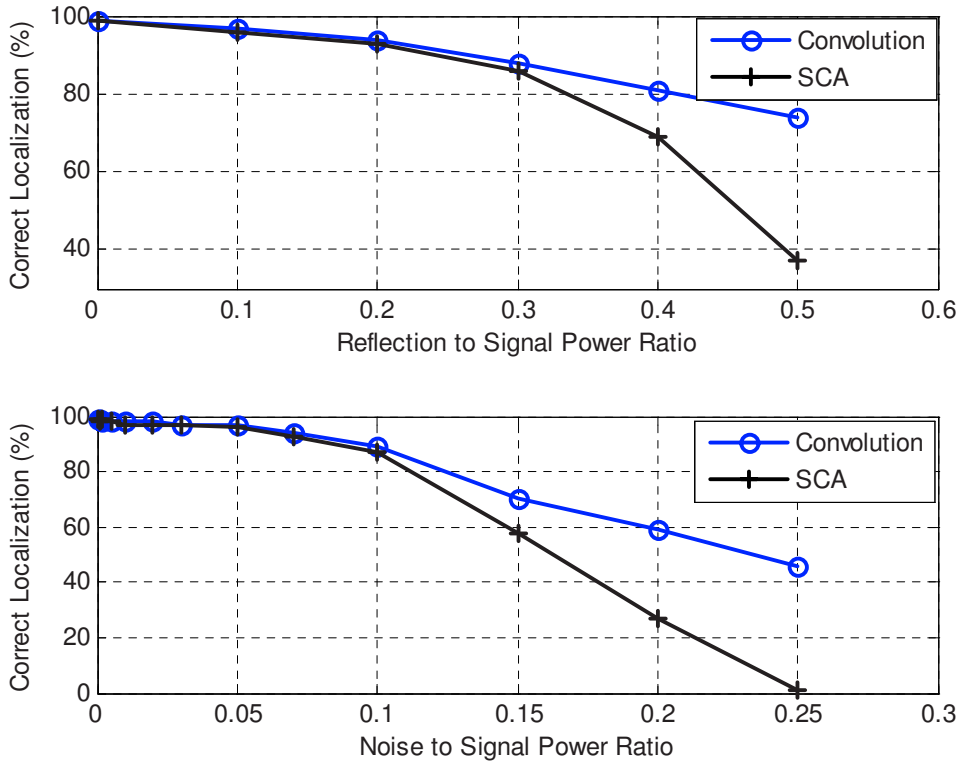


Figure 4.10: Comparison of cross convolution method and SCA in presence of reflections and noise.

which front-back reversals in the azimuthal plane increase and the rate of erroneous detections increases. To correct these false localization errors and decrease the computational complexity, we shall apply an extended Kalman filtering to the output of the convolution based localizer.

### 4.2.1 Extended Kalman Filtering

The cross convolution based algorithm described in the previous section yields a unique azimuth and elevation corresponding to a specific HRTF pair. This azimuth will constitute the input to the extended Kalman filter [18] properly tuned to minimize the front-back reversals in the azimuthal plane. The Kalman filter will update and track the sound source on a 'per scan' basis where 'scan' means an interval of time after which an azimuth estimate is produced by the convolution based algorithm. In order to remove front-back azimuth reversals and false detections in the azimuthal plane,we have used horizontal turn model which can track the sound source in the azimuthal plane. The state vector $\mathbf{x}$ is defined as $\mathbf{x} = [x, v_x, y, v_y]^T$, where $x$ and $y$ are the sound source coordinates,$v_x$ and $v_y$ are the corresponding

velocities [18]. The turning rate $w$ of the sound source is updated in every scan. If the difference of azimuths, estimated using the convolution based localization algorithm, in the current and previous scans is beyond an allowable maximum value of 100°, we do not update the turning rate and we check for a possible occurrence of front-back reversal. In case we have an erroneous detection without azimuth reversal, we keep the turning rate unchanged and ignore the azimuth in the current scan. The azimuth estimated by the Kalman filter will be equal to azimuth predicted in the last scan. For horizontal turn model, the state transition matrix $\Phi$ is defined as,

$$\Phi = \begin{bmatrix} 1 & sinwT/w & 0 & -(1-coswT)/w \\ 0 & coswT & 0 & -sinwT \\ 0 & (1-coswT)/w & 1 & sinwT/w \\ 0 & sinwT & 0 & coswT \end{bmatrix} \tag{4.7}$$

where T is the scan time.
The process noise covariance is given as

$$\mathbf{Q} = \sigma_{q^2} \begin{bmatrix} T^4/4 & T^3/2 & 0 & 0 \\ T^3/2 & T^2 & 0 & 0 \\ 0 & 0 & T^4/4 & T^3/2 \\ 0 & 0 & T^3/2 & T^2 \end{bmatrix} \tag{4.8}$$

where $\sigma_{q^2}$ is the random acceleration variance. The matrix $\mathbf{Q}$ represents random motion entering the system between sampling intervals.
We have defined our measurement vector as

$$\mathbf{z}(k) = [rcos\theta(k)cos\varphi(k), rsin\theta(k)cos\varphi(k)]^T \tag{4.9}$$

where r,$\theta$ and $\varphi$ are the source range,azimuth and elevation.The measurement matrix is given as,

$$\mathbf{H_x}(k) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{4.10}$$

Since the sound source dynamics are difficult to be predicted with single model, we use an interactive multiple model scheme [18] to track sound in the azimuthal plane.

### 4.2.2 Implementation

Figure 4.11 shows the block diagram for real time implementation of HRTF-Kalman algorithm. Using two microphones inserted in the ear canals of a humanoid head, the data acquisition module acquires 10000 samples per second from a moving sound source. A moving average filter is used to smooth the acquired data. The convolution

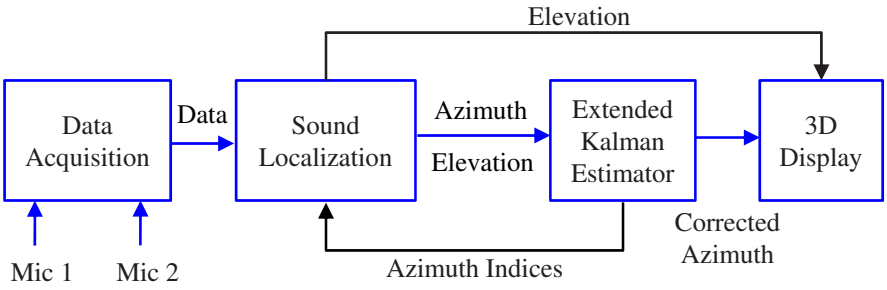


Figure 4.11: Block diagram of the HRTF-Kalman algorithm.

based HRTF algorithm acquires the data from the microphones inside the ear canal of the humanoid head as explained in previous sections, and estimates the azimuth and elevation of the sound source. The estimated azimuth constitutes the input to the extended Kalman filter which is responsible for correcting azimuthal localization error caused by front-back reversal. The estimated azimuth and elevation are sent out to a 3D display. In the first scan, the convolution based HRTF algorithm has to run over all possible azimuths to check for maximum correlation. Depending on the association of the sound source dynamics in the current and previous scans, a track will be initiated. After a track is initiated, a region of interest is created and the localization algorithm has access to a selected set of HRTF pairs corresponding to indices of azimuths in the vicinity of the azimuth that was predicted in the last scan. Thus, the computational burden of convolving with all HRTF pairs is considerably minimized and the localization speed increased by a factor of 85% on average.

### 4.2.3 Performance Analysis

In our experimental setup, we have placed a KEMAR artificial torso and head equiped with two silicon ears in a highly reverberant room exhibiting reflections and background noise like computer fans and people walking. The different parameters for the extended Kalman filter are set to the following numerical values, $\sigma_q$=0.1 $m/sec^2$, $\sigma_\nu$=0.1 m, $T$ =0.5sec, where $\sigma_q$ and $\sigma_\nu$ are the random acceleration and measurement standard deviations, $T$ is the scan time. Different audio signals including male and female speech signals as well as broadband clicks were moved simultaneously around the humanoid head at different elevation angles. Both convolution based and HRTF-Kalman algorithms are set to process the data collected by the two microphones inside the ear canals of the humanoid. Using the convolution based method without Kalman processing, the front-back reversal problem emerges and 74% of the estimated azimuth and elevation are detected correctly at there target location. Nevertheless, deploying the HRTF-Kalman approach under

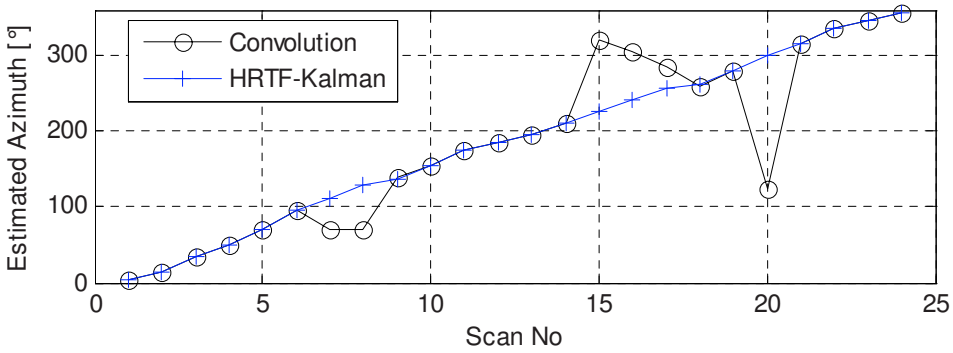


Figure 4.12: Localization performance averaged over different speech and broadband sound signals moving clockwise around the head.

the same high reverberation condition, the average percentage of correct localization improved to 91% and the front-back reversal problem is resolved. Figure 4.12 shows the results of both algorithms for a male speech source moving around the head from 0° to 360° in azimuth. Using only convolution based HRTF localization algorithm, several front-back azimuth reversals occur along with occasional erroneous detections. For example, a front-back reversal error is observed when the convolution based localization algorithm yields azimuth values of 70° and 305° for sound source actually located at 110° and 235°, where azimuth 0° is defined as the location directly facing the humanoid. This problem is completely resolved with the help of extended Kalman filter and the overall performance increases from 73% to 94% of correctly localized sound sources. The remaining 6% deviate 1.7° on average from their target location. This is due to the differences between the dummy manikin model used in the experiment and the KEMAR model used to obtain the HRTF dataset.

Furthermore, we compare our method with the robotic sound localization system proposed in [49]. Within this framework, 4 microphones are appropriately arranged on the surface of a spherical head, three at the side and one on the top. The microphones were placed 15 cm apart, i.e. about 1.5 times the interaural distance of humans. A sphere-shaped head was used to simplify the formulation of the time difference calculation. By using the top-mounted microphone, the elevation of sound sources based on the time difference can be localized without using the relatively uncertain spectral difference cue.

Similar to [49], we have conducted sound localization experiments with testing sounds including coin-dropping, glass-broken and a piece of classic music were used. The sampling frequency was 44,1kHz. The distance of the sound source was set to 1.0m. Elevation testing was performed for -15, 0, 15, 30, 60 and 90 degrees, with the azimuth set to 0 degree. Azimuth testing was performed for -90, -45, 0,

Table 4.2: Performance comparison with the arrival time difference method [49].

| **Number of microphone pairs** | **6** | **5** | **4** | **3** | **2** | **Cross Convolution** |
|---|---|---|---|---|---|---|
| **Azimuth** | 2.1° | 2.1° | 2.4° | 1.5° | 2.6° | 1.9° |
| **Elevation** | 2.9° | 2.9° | 2.7° | 2.6° | 2.7° | 2.3° |

45, 90 degrees with the elevation set to low (-15 degrees), mid (0 degree) and high (60 degrees) for all cases. The average localization errors using different number of microphone pairs are shown in Table 4.2. It is shown that choosing three microphone pairs with smallest arrival time difference achieved the best performance.

Compared to the method in [49], our cross convolution algorithm deploys two, instead of four, microphones and is outperforming in elevation while exhibiting a similar localization performance in azimuth. It should be noted, however, that the average errors revealed by the arrival time difference method in Table 4.2 are for sound sources located in the frontal hemisphere with respect to the spherical head, i.e. for azimuths between -90 and 90 degrees. The performance of the method degrades considerably for sound sources coming from behind the spherical robot head, while this is not the case for the SCA algorithm.

Applying the presented sound localization methods, chapter 5 addresses the challenging task of the concurrent sound source localization and separation in reverberant environments, using only two microphones.

# Chapter 5

# Concurrent Sound Source Localization and Separation

We combine binaural sound-source localization and separation techniques for an effective deployment in telerobotics. Relying on the concept of binaural hearing, where the human auditory 3D percepts are predominantly formed on the basis of the sound-pressure signals at the two eardrums, our robotic 3D localization system uses only two microphones placed inside the ear canals of a robot head equipped with artificial ears and mounted on a torso. The challenging task of using only two microphones for 3D localization is made more intriguing by allowing more sources to coexist and randomly move within the robot's environment. The proposed localization algorithm exploits all the binaural cues encapsulated within the HRTFs. Taking advantage of the sparse representations of the ear input signals, the 3D positions of three concurrent sound sources are extracted. The location of the sources is extracted after identifying which HRTFs they have been filtered with using a well-known self-splitting competitive learning clustering algorithm. Once the locations of the sources are identified, they are separated using a generic HRTF dataset. Simulation results demonstrated highly accurate 3D localization of the two concurrent sound sources, and a very high Signal-to-Interference Ratio (SIR) for the separated sound signals.

## 5.1 Motivation

One of the most challenging characteristics of human spatial hearing, is the cocktail party phenomenon, where attention pertains to the ability of a listener to focus on one channel while ignoring other irrelevant channels. In robotics, on the other hand, efficient and accurate binaural 3D localization of several sound sources is quite a challenging task. In recent years, a good number of algorithms have been proposed to tackle this problem. Basically, most of the detection methods used rely on microphone arrays, where the number of microphones is more or equal to the number of sound sources to be localized concurrently in 3D [52]. Among them, some approaches deal with simultaneous localization and separation of sound sources [123]. However, a more intriguing, and naturally more demanding scenario, is localization of sound sources that outnumber the available number of microphones. Very few approaches were able to estimate the position of the sound sources, while only providing azimuth angles [94]. Humans and most mammals, however, are capable of detecting multiple concurrent sound sources with two ears by assessing monaural cues and binaural cues like ILD and ITD, in several frequency bands.

Recent investigations on the auditory space map of the barn owl, a predator with an astonishing ability to localize sound, revealed that the ILD/ITD cues cluster around two positions in the auditory map when two uncorrelated sound sources are simultaneously present. These clusters stem from time-frequency instances when one source predominates the other, i.e. has a stronger intensity [57]. Using this fact we present two new approaches for separating and localizing two or more concurrent sound sources in 3D using only two small microphones placed inside the KEMAR artificial ears.

Section 5.2 describes an algorithm for two sound sources that iteratively adapts the coefficients of a MIMO system and provides the two statistically independent source signals. This well-known separation method which exploits the non-stationarity of the sources is used to retrieve two speakers from two convolutive mixtures. By using a simple relation between blind source separation and system identification, the HRTFs that filtered the sound sources can be determined under the condition of an anechoic environment.

The second algorithm, presented in section 5.3, applies Short-Time Fourier Transform (STFT) to the ear signals and makes use of the sparseness of the sources in time-frequency domain. If the concurrent signals are sparse, which is naturally the case with speech signals, there must exist many instances when one source predominates the other. In such cases, the ear signals cluster around the actual HRTF, corresponding to the correct source location, in the single frequency bands. The positions of the sources are finally determined by a database lookup. With the

respective HRTFs of the database, the sources can be separated by inversion of the HRTF system in case of two concurrent sound sources or by L1-norm minimization in case of more than two sources. The CIPIC HRTF database containing HRIRs for 1250 positions in 3D space was used for the database lookup [6]. Its HRIRs were measured at the CIPIC Interface Laboratory at the university of California, Davis. The simulation results in section 5.5 show that both algorithms localize two and more concurrent sound sources with a high accuracy and separate them with little cross-talk in the output signals [70].

## 5.2 Source Separation and Localization Using Adaptive MIMO-Filtering

### 5.2.1 Source Separation

One can consider the space between the sound sources and the microphones of the artificial head with its different transmission paths as a multiple-input multiple-output (MIMO) system, with the two sound sources as the input signals and the two ear signals as the output signals [19]. This system can be described by

$$x_1 = h_{11} \cdot s_1 + h_{21} \cdot s_2 \tag{5.1}$$

$$x_2 = h_{12} \cdot s_1 + h_{22} \cdot s_2 \tag{5.2}$$

where $s_1$ and $s_2$ represent the two concurrent sound sources, $h_{11}$ and $h_{22}$ correspond to the HRTFs for the direct paths from the two sources to the head, $h_{12}$ and $h_{21}$ represent the crossing channels, and $x_1$, $x_2$ model the mixed signals collected by the microphones at the end of the ear canals. Figure 5.1 illustrates the differing transmission paths from the two sound sources to the eardrums. These paths may be regarded as linear, time-invariant systems and are represented by the HRTFs in (5.1) and (5.2). The aim is now to find out the HRTFs the source signals were filtered with on their way to the microphones. This is normally done by multichannel blind deconvolution (MCBD) methods. However, the problem with traditional MCBD methods is that the output signals are temporally whitened. To avoid this, we use a convolutive blind source separation (BSS) algorithm proposed in [24] which prevents such whitening effects. Its major advantage is that, if it is implemented in an efficient way, it can be used for real-time applications. We perform BSS in the time-domain in order to avoid the permutation problem which arises in combination with frequency-domain convolutive independent component analysis in each frequency bin. Figure 5.2 shows the demixing filter system forcing its output signals $y_1(k)$ and $y_2(k)$ to be mutually statistically independent. In order to achieve this, we use second-order statistics which exploit the following two properties of the source signals: 1)

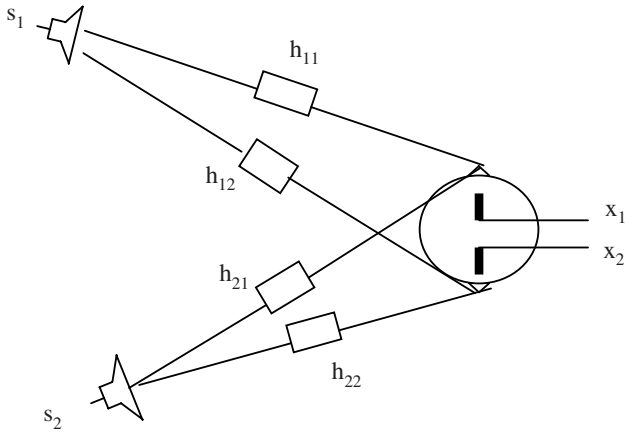


Figure 5.1: Differing transmission paths between sound sources and artificial head microphones modeled as a HRTF MIMO system.

Non-whiteness, by simultaneous diagonalization of output correlation matrices over multiple time-lags, 2) Non-stationarity, by simultaneous diagonalization of short-time output correlation matrices at different time intervals. The proposed algorithm is a block-online algorithm, i.e. it processes subsequent blocks from the microphone mixture signals over a certain number of iterations in the off-line mode in order to train the demixing filters. In [24] the cost function

$$J(m) = \sum_{i=0}^{m} \beta(i,m)\{\log\det bdiag Y^H(i)Y(i) - \log\det Y^H(i)Y(i)\} \tag{5.3}$$

is proposed for source separation, where $\beta$ is a window function that is normalized according to $\sum_{i=0}^{m}\beta(i,m) = 1$. This is meant to allow tracking in time-varying environments (moving sound sources). The *bdiag* operation applied on a partitioned block matrix consisting of several sub-matrices sets all sub-matrices on the off-diagonals to zero. The parameter $m$ is a block time index. The output signal matrix $Y$ can be written as

$$Y(m) = \begin{bmatrix} Y_1(m) & Y_2(m) \end{bmatrix} \tag{5.4}$$

$$Y_q(m) = \begin{bmatrix} Y_q(mL) & \dots & Y_q(mL-L+1) \\ Y_q(mL+1) & \ddots & Y_q(mL-L+2) \\ \dots & \ddots & \vdots \\ Y_q(mL+N-1) & \dots & Y_q(mL-L+N) \end{bmatrix} \tag{5.5}$$

where $L$ denotes here the length of the demixing filters and $N$ is the block length. The $N \times L$ matrix $Y_q(m)$ incorporates $L$ time-lags in the correlation matrices into the

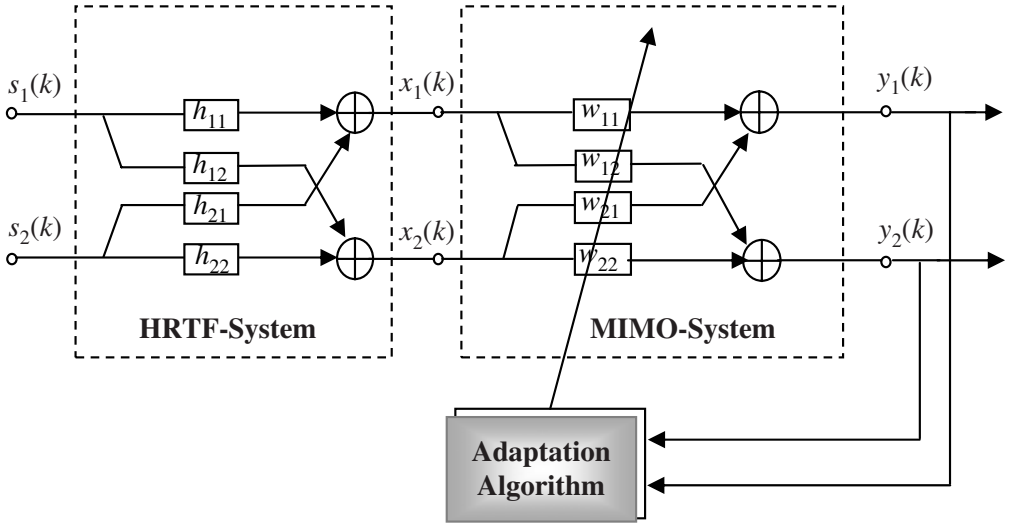


Figure 5.2: Demixing system with filters $w_{11}$, $w_{12}$, $w_{21}$ and $w_{22}$ which are adapted by simultaneous diagonalization of short-time correlation matrices of output signals $y_1(k)$ and $y_2(k)$ in order to separate the sound sources.

cost function in (5.3), which is necessary for the exploitation of the non-whiteness property. Applying the natural gradient algorithm the updates

$$\Delta W(m) = 4 \sum_{i=0}^{m} \beta(i,m) W \cdot \begin{bmatrix} 0 & \dots & R_{y_1 y_2} R_{y_2 y_2}^{-1} \\ R_{y_2 y_1} R_{y_1 y_1}^{-1} & \ddots & 0 \end{bmatrix} \tag{5.6}$$

are obtained for the coefficients of the demixing filters [24], where $W$ is a $4L \times 2L$ matrix containing the filter coefficients of $w_{11}$, $w_{12}$, $w_{21}$ and $w_{22}$. The matrix $W$ has the Sylvester structure

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \tag{5.7}$$

$$W_{pq}(m) = \begin{bmatrix} w_{pq,0} & 0 & \dots & 0 \\ w_{pq,1} & w_{pq,0} & \ddots & \vdots \\ \vdots & w_{pq,1} & \ddots & 0 \\ w_{pq,i-1} & \vdots & \ddots & w_{pq,0} \\ 0 & w_{pq,i-1} & \ddots & w_{pq,1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & w_{pq,i-1} \\ 0 & \dots & 0 & 0 \end{bmatrix} \tag{5.8}$$

Note that the update matrix $\Delta W$ is also a $4L \times 2L$ matrix. It should also be noted that if, for every iteration of the algorithm, the whole matrix $\Delta W$ were to be calculated, the computational complexity of the algorithm would become very high. Observing that all the coefficients of the demixing filters are in the first columns of the matrices $W_{pq}$, it is not necessary to compute all the column entries to update the filters. Thus, only the first column entries of the cross-correlation matrices $R_{y_1y_2}$ and $R_{y_2y_1}$ denoted by the following column vectors have to be calculated.

$$\mathbf{corr}_{x_py_q} = \begin{bmatrix} r_{y_py_q}(m,0) & r_{y_py_q}(m,-1) & \dots & r_{y_py_q}(m,-L+1) \end{bmatrix}^T \tag{5.9}$$

$$r_{y_py_q}(m,k) = \left[\sum_{n=mL+|k|}^{mL+N-k-1} y_p(n+k)y_q(n)\right]^T \qquad k \leq 0 \tag{5.10}$$

The computation of the inverses of the autocorrelation matrices $R_{y_qy_q}(m)$ can also be simplified. As the separated signals $y_1(k)$ and $y_2(k)$ are assumed to be stationary within each block, we can approximate the autocorrelation matrices by

$$R_{y_qy_q}(m) \approx \left(\sum_{n=mL}^{mL+N-k-1} y_q^2(n)\right) I = \sigma_{yq}^2(m) I \tag{5.11}$$

which leads to an element-wise division of the correlation vectors $\mathbf{corr}_{y_py_q}(m)$ by the output signal energy. Hence, the updates for the demixing filters result in

$$\Delta W_\gamma(m) = 4 \sum_{i=0}^{m} \beta(i,m) W \cdot \begin{bmatrix} 0 & \frac{\mathbf{corr}_{y_1y_2}}{\sigma_{y_2}^2} \\ \frac{\mathbf{corr}_{y_2y_1}}{\sigma_{y_1}^2} & 0 \end{bmatrix} \tag{5.12}$$

As $W$ has Sylvester structure, the matrix product in (5.12) can be implemented as a convolution of the cross-correlation vectors with the impulse responses of the demixing filters. The new filter coefficients are finally calculated with the update equation

$$W_r = W_r(m-1) - \mu \Delta W_r(m) \tag{5.13}$$

where $\mu$ denotes the stepsize. The index $r$ refers to the reduced $2L \times 2$ matrices. As already mentioned, the algorithm we are using is a block-online algorithm which acquires $KL+N$ samples from the two microphone mixture signals $x_1(k)$ and $x_2(k)$ and divides them into $K$ off-line blocks which are simultaneously processed for a certain number of iterations. The microphone signals of each off-line block are convolved with the demixing filters $w_{pq}$ of the previous iteration in order to get the output signals $y_1(k)$ and $y_2(k)$. After calculating the signal energies $\sigma_{yq}^2$, and the cross-correlation vectors $\mathbf{corr}_{y_py_q}(m)$ of each off-line block, the matrix product in (5.12) is calculated. Afterwards, the $K$ matrix products are averaged and the filters are updated for the next iteration using (5.13). At the end of all iterations, the off-line part is finished and the overall update is calculated, taking into account the

window function $\beta$. Finally, the new demixing filters are used to process the next on-line block that consists of $KL + N$ new samples. A detailed pseudo-code of this algorithm can be found in [3].

In order to ensure robust convergence, an appropriate choice of the stepsize $\mu$ is important. According to [25], we employed an adaptive stepsize that is calculated before each coefficient update by

$$\mu(m+1) = \begin{cases} 1.1 \cdot \mu(m) & J(m) < J(m-1) \\ 0.5 \cdot \mu(m) & J(m) \geq 1.3 \cdot J(m-1) \\ \mu(m) & \text{otherwise} \end{cases} \tag{5.14}$$

Furthermore, to avoid instabilities, the range of the stepsize has to be restricted to $[_{min}, _{max}]$. According to (5.14), the cost function $J(m)$ has to be evaluated in each iteration step. This, however, would result in a high computational complexity since the matrix product $Y^H Y$ is required, see (5.3). In order to increase efficiency, we just take the $L_2$-norm of the cross-correlation vectors $\mathbf{corr}_{y_p y_q}(m)$ that were already calculated in a previous step and consider it as an appropriate substitute for $J(m)$.

### 5.2.2 System Identification

After source separation and in order to determine the positions of the two sound sources in azimuth and elevation, we have to identify the HRTFs with our adaptive demixing filters $w_{11}$, $w_{12}$, $w_{21}$ and $w_{22}$. The overall MIMO system of Figure 5.2 with the two source signals $s_1(k)$ and $s_2(k)$ as inputs and the separated signals $y_1(k)$ and $y_2(k)$ as outputs is the concatenation of the HRTF system and the MIMO system containing the demixing filters. Its direct paths and cross paths are illustrated in Figure 5.3. Obviously, to ideally separate the sources, its cross-channels from the first source to the second ear and from the second source to the first ear, denoted as $c_{12}$ and $c_{21}$, must be forced to zero. This can be expressed by

$$c_{12} = h_{11} \cdot w_{12} + h_{12} \cdot w_{22} = 0 \tag{5.15}$$

$$c_{21} = h_{21} \cdot w_{11} + h_{22} \cdot w_{21} = 0 \tag{5.16}$$

The correct HRTFs are found similarly to the method described in chapter 4, i.e. all possible HRTFs of the KEMAR database for elevation angles between -40 and 90 and azimuth angles between 0 and 355 are convolved with the corresponding demixing filters according to (5.15) and (5.15). The pairs that yield minimum norms of the vectors $c_{12}$ and $c_{21}$ are considered to be the correct HRTFs that originally filtered the sound sources. In practice, the two equations can be evaluated independently of each other. Equations (5.15) and (5.16) provide the position of the two concurrent

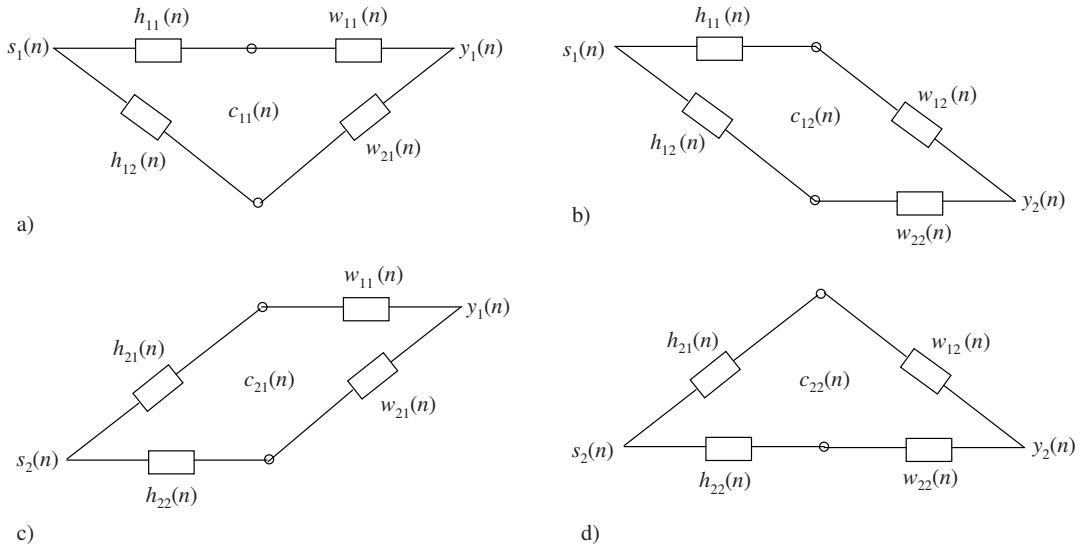


Figure 5.3: The combination of HRTF system and MIMO system for blind source separation can be divided into four SIMO-MISO systems.

sources. The combination of the two processes of blind source separation and system identification is done in parallel for fast tracking of moving sound sources as shown in Figure 5.4. Once the source separation algorithm has converged, the system identification process can use the adapted filters to determine the correct HRTFs. Meanwhile, the source separation process can take new on-line blocks for further coefficient updates. After the identification process has finished it can take the updated filters to determine the new positions and so forth [79].

## 5.3 Localization and Separation by Clustering in Time-Frequency Domain

### 5.3.1 Blind System Identification Framework

A second approach to binaural sound localization is based on finding the HRTFs, the sound sources were filtered with, on their way to the robot's microphones, which, in our case, play the role of the human eardrums. Applying Short-Time Fourier Transform (STFT), we can describe the ear input signals with the following equations:

$$X_1(f,\tau) = \sum_{j=1}^{M} H_{1j}(f) S_j(f,\tau) \tag{5.17}$$

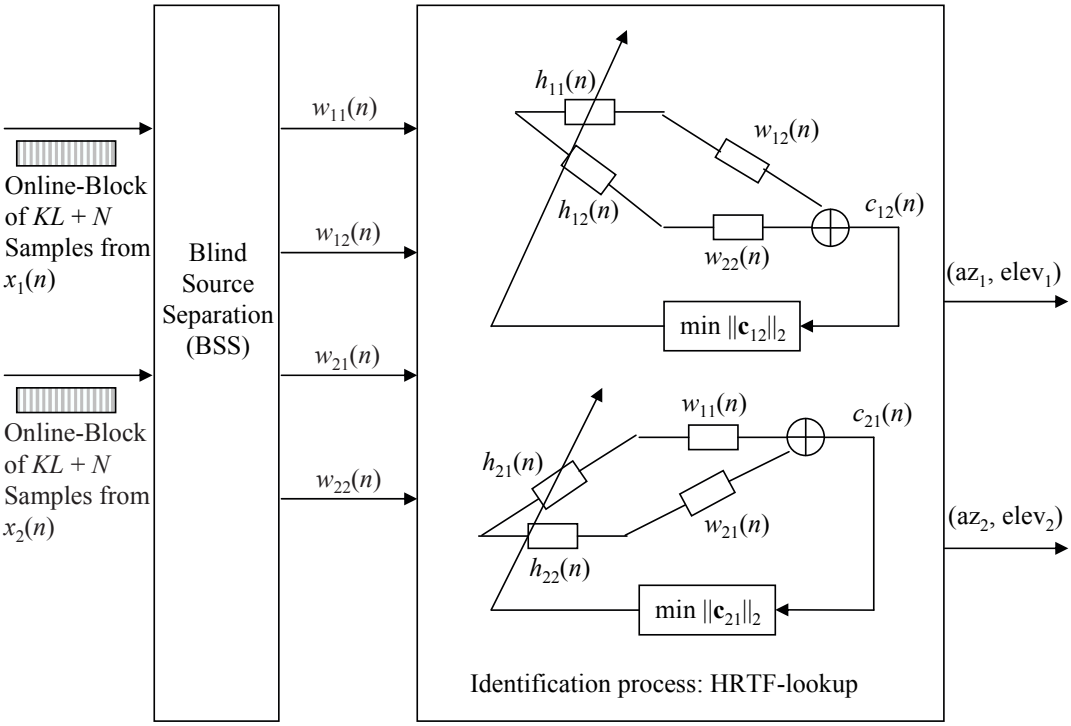


Figure 5.4: Simultaneously running processes: BSS and HRTF lookup. The HRTF lookup process takes the adapted filters $w_{11}$, $w_{12}$, $w_{21}$ and $w_{22}$ from the BSS process and finds the azimuth and elevation positions of the first source (az1, elev1) and of the second source (az2, elev2).

$$X_2(f,\tau) = \sum_{j=1}^{M} H_{2j}(f) S_j(f,\tau) \tag{5.18}$$

where $\tau$ denotes the time frame. The term $M$ is the number of sound sources, and $S_j(f,\tau)$ are windowed sound source signals in frequency-domain. It is known that speech signals are very sparse in time-frequency domain, more than in time-domain [138]. However, frequency domain Independent Component Analysis (ICA) introduces the inherent permutation problem in each frequency bin, to which we will later present a solution. Since a sparse signal is almost zero in most time-frequency instances, there are a many instances when only one source is active. Hence, the ear input signals can be rewritten as:

$$X_1(f,\tau) = H_{1J}(f) S_J(f,\tau) \tag{5.19}$$

$$X_2(f,\tau) = H_{2J}(f) S_J(f,\tau) \quad J \in \{1,\ldots,M\} \tag{5.20}$$

Assuming stationary source positions, the HRTFs $H_{1J}(f)$ and $H_{2J}(f)$ are constant for all time instances $\tau$. Since they are related to the source positions they are

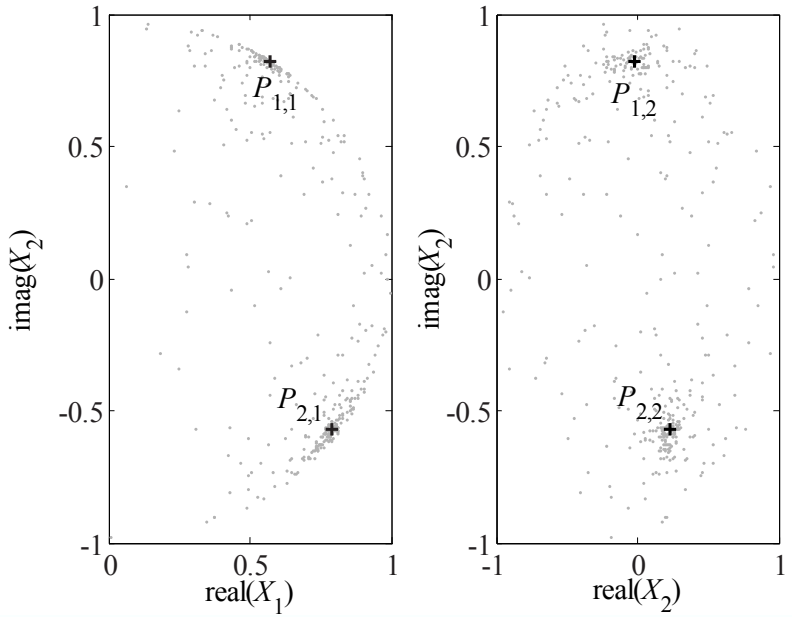


Figure 5.5: Samples from the two ear microphones after STFT. The two subplots depict the real part of $X_1$ and the real part of $X_2$ versus the imaginary part of $X_2$, respectively. The data was gathered from 400 time-frames at a frequency of 538 Hz and normalized according to (5.21) and (5.22). The clusters show that there are two speakers present. Furthermore, the prototypes determined by Self-Splitting Competitive Learning are depicted in the cluster centers.

different for each source. This means ideally, that the time-frequency samples of the ear input signals, $X_1(f,\tau)$ and $X_2(f,\tau)$ , that originate from the $J$-th source, cluster at each frequency $f$ around the corresponding complex HRTFs values. Additionally, the Fourier transforms of the ear input signals are phase and amplitude normalized:

$$X_1(f,\tau) \leftarrow \frac{X_1(f,\tau)}{\sqrt{X_1^2(f,\tau) + X_2^2(f,\tau)}} \exp^{-\varphi_{x_1}} \tag{5.21}$$

$$X_2(f,\tau) \leftarrow \frac{X_2(f,\tau)}{\sqrt{X_1^2(f,\tau) + X_2^2(f,\tau)}} \exp^{-\varphi_{x_1}} \tag{5.22}$$

where $\varphi_{x_1}$ is the phase corresponding to the input signal of the left ear microphone, which is chosen as a reference sensor. Figure 5.5 illustrates the clustering of the normalized data in the feature space.

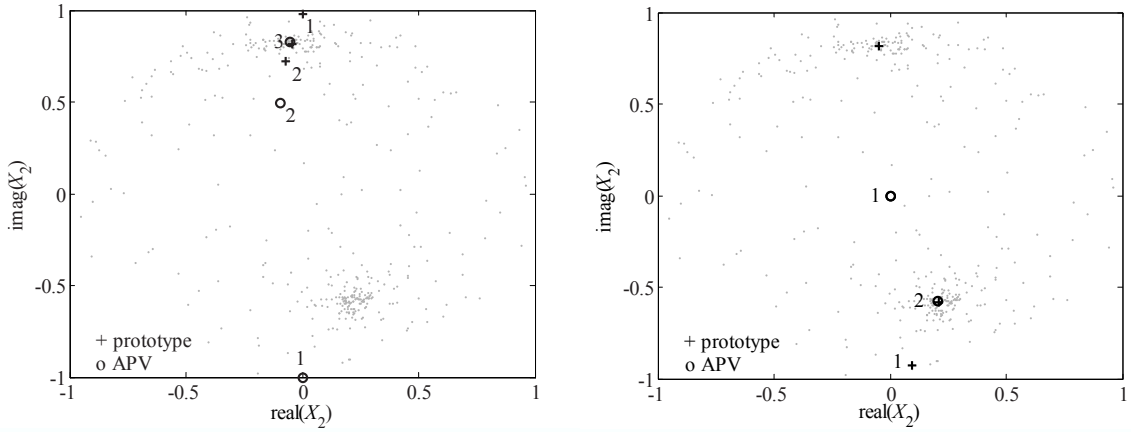


Figure 5.6: **Left**: Learning process of the first prototype. Step 1 shows the initialization of the second component of $\vec{P_1}$ and the APV $\vec{A_1}$. Step 2 shows their positions after 100 iterations. In step 3 the distance between $\vec{P_1}$ and $\vec{A_1}$ has fallen to 0.01 and learning stops. **Right**: Learning of prototype $\vec{P_2}$ created after $\vec{P_1}$ has settled in the center of the topmost cluster. It is initialized together with an APV $\vec{A_2}$ at a certain distance from the first prototype (step 1), and is led to the center of the cluster at the bottom after some 100 iterations (step 2).

### 5.3.2 Self-Splitting Competitive Learning

As pointed out earlier, the source separation problem needs to be solved in each frequency bin. This means that, our algorithm clusters the data in all frequency bins over several time frames separately. For further data analysis we use the clustering algorithm proposed in [140] which is based on self-splitting competitive learning (SSCL). In the following, we will briefly describe its principle. The key issue in SSCL is the One-Prototype-Take-One-Cluster paradigm (OPTOC). this means that one prototype represents only one cluster. At first, a single prototype $\vec{P_1} = \begin{bmatrix} P_1 & P_2 \end{bmatrix}^T, (P_1, P_2 \in \mathbb{C})$ is initialized randomly in the feature space. At the same time an asymptotic property vector (APV), $\vec{A_1} = \begin{bmatrix} A_1 & A_2 \end{bmatrix}^T, (A_1, A_2 \in \mathbb{C})$, is created far away from the prototype. Its task is to guide the learning of the prototype making sure that, after some iterations, the prototype has settled in the center of a cluster (Figure 5.6).

The update of $\vec{A_1}$ in each iteration is calculated as

$$\vec{A_1^\star} = \vec{A_1} + \frac{1}{n_A} \cdot \delta_1 \cdot (\vec{X} - \vec{A_1}) \cdot \Phi(\vec{P_1}, \vec{A_1}, \vec{X}) \tag{5.23}$$

$$\Phi(\vec{P_1}, \vec{A_1}, \vec{X}) = \begin{cases} 1 & ||\vec{P_1} - \vec{A_1}|| \geq ||\vec{P_1} - \vec{X}|| \\ 0 & \text{otherwise} \end{cases}$$

where $||\cdot||$ is the Euclidean norm. The elements of $\vec{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^T$, $(X_1, X_2 \in \mathbb{C})$ are randomly chosen patterns of the normalized data in (5.21) and (5.22) at a certain time-frequency instant. The term $\delta_1$ can be set constant or it can be calculated as

$$\delta_1 = \left( \frac{||\vec{P_1} - \vec{A_1}||}{||\vec{P_1} - \vec{X}|| + ||\vec{P_1} - \vec{A_1}||} \right)^2 \tag{5.24}$$

The winning counter $n_{A_1}$ is updated as

$$n_{A_1} = n_{A_1} + \delta_1 \cdot \Phi(\vec{P_1}, \vec{A_1}, \vec{X}) \tag{5.25}$$

The APV $\vec{A_1}$ thus defines a neighborhood around the prototype $\vec{P_1}$. If a randomly taken pattern, $\vec{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^T$, obtained using (5.21) and (5.22) lies within this neighborhood, it contributes to learning $\vec{A_1}$. It is observed that in the course of iterations, $\vec{A_1}$ moves towards $\vec{P_1}$. Learning stops when the Euclidean norm $|||\vec{P_1} - \vec{A_1}|||$ falls below a constant $\epsilon_1$.

Now, in order to classify other clusters that may be present in the feature space, further prototypes have to be initialized. Hence, the following split validity criterion is introduced. If $|||\vec{P_1} - \vec{A_1}|||$ is smaller than a constant $\epsilon$, a new prototype and a new APV are created in the feature space which are to lead to the center of another cluster (Figure 5.6). The learning process starts anew. The term $\vec{C_1}$ is called the Center Property Vector (CPV) and determines the arithmetic mean of the input data points which have contributed to learning the prototype $\vec{P_1}$ . In order to avoid unnecessary competition between the first and the new prototype, a distant property vector $\vec{R_1}$, adapted during the learning process, makes sure that the new prototype $\vec{P_2}$ is initialized far away from the first one. A detailed pseudo-code of the SSCL algorithm and update equations are given in [140].

Finally, the adaptation of the $i$-th prototype proceeds according to the following equation

$$\vec{P_i^\star} = \vec{P_i} + \alpha_i \cdot \beta_i \cdot (\vec{X} - \vec{P_i}) \tag{5.26}$$

where

$$\alpha_i = \left( \frac{||\vec{P_i} - \vec{A_i}||}{||\vec{P_i} - \vec{X}|| + ||\vec{P_i} - \vec{A_i}||} \right)^2 \tag{5.27}$$

$$\beta_i = \left( \frac{||\vec{P_i} - \vec{R_i}||}{||\vec{P_i} - \vec{X}|| + ||\vec{P_i} - \vec{R_i}||} \right)^2 \tag{5.28}$$

The equations (5.23) - (5.28) also hold for the learning of the $i$-th APV and prototype if the index 1 is changed to $i$. A detailed pseudo code of the self-splitting competitive learning algorithm is given in [140].

A crucial key directly affecting the performance of the clustering algorithm is the choice of the two constants $\epsilon_1$ and $\epsilon_2$. As opposed to an adaptive choice, proposed in [138], that depends on the variances and number of elements of the clusters, we set and to a constant value, 0.01 in our case, and we confine the maximum number of prototypes to the number of present sound sources plus two. On the one hand, this has the disadvantage that the algorithm does not work completely blindly as the number of present sources has to be known but, on the other hand, it results in a robust classification of the clusters. The maximum number of clusters is chosen a little bit larger than the actual number of sources, since there are many data points, in the feature space, resulting from non-sparse time-frequency instances, see Figure 5.6. These data points should not be represented by the prototypes in the center of the clusters, but by other prototypes. Of course, there has to be a criterion in order to choose the "right" prototypes that represent the HRTFs at a certain frequency.

A criterion that proved to be appropriate goes as follows:

$$If \quad \frac{K_i}{v_i} > \frac{1}{4} \underset{i}{max}\big(\frac{K_i}{v_i}\big) \quad \forall i = 1, \ldots, \text{number of prototypes},$$

then the $i$-th prototype represents a HRTF at a certain frequency. Hereby, $K_i$ denotes the number of data points that have been assigned to the $i$-th prototype, and $v_i$ is the variance of the cluster $S_i$ which is a-posteriori determined by

$$v_i = \frac{1}{K_i - 1} \sum_{X \in S_i} ||\vec{X} - \vec{P_i}|| \tag{5.29}$$

Figure 5.7 shows five prototypes at the end of the learning process in case of three concurrent speakers. Note that the prototypes $P_4$ and $P_5$ do not fulfill the criterion above and are therefore discarded.

### 5.3.3 Solving the Permutation Problem

As mentioned earlier, we have to tackle the permutation problem introduced by frequency-domain independent component analysis. Once the clusters in the feature space in all frequency bins have been classified, one has to determine which of the clusters in the frequency bins belong to the same HRTF. Our approach is based on the assumption that the position of the clusters does not move a lot between adjacent frequency bins. The prototypes that remain, after applying the above-mentioned criterion, in the frequency bin $f$ can be arranged in a matrix $H(f) = \big[\vec{P_1}(f)\vec{P_1}(f) \ldots \vec{P_{N_{Pr}}}(f)\big] \in \mathbb{C}^{2 \times N_{Pr}}$. The variable $N_{Pr}$ denotes the number of remaining prototypes that represent the HRTFs. Let $\Psi = \{\Pi_1, \Pi_2, \ldots, \Pi_M\}$ be a group of permutation matrices of dimension $M \times M$. Then, the correct permutation

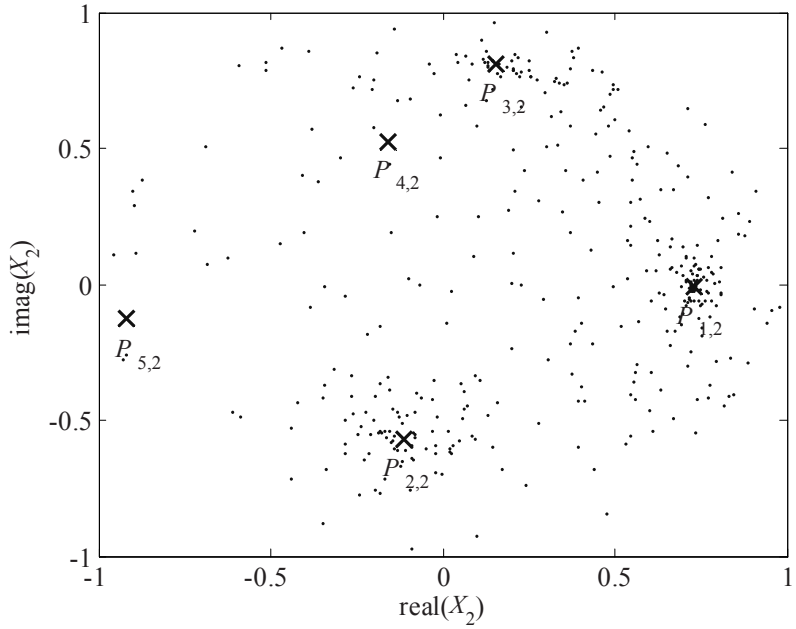


Figure 5.7: Five prototypes at the end of the learning process with three concurrent sound sources. This figure shows imaginary and real parts of the second component of the prototypes. In this frequency bin, only $\vec{P}_1$, $\vec{P}_2$ and $\vec{P}_3$ represent the HRTFs which filtered the sound sources. $\vec{P}_4$ and $\vec{P}_5$ are discarded.

can be described by

$$\begin{bmatrix} d_{i1} & d_{i2} & \dots & d_{iN_{Pr}} \end{bmatrix} = H(f) \cdot \Pi_i^T - H(f-1) \quad \forall \Pi_i \in \Psi \tag{5.30}$$

$$\hat{\Pi} = \Pi_i \cdot \min_i \sum_{j=1}^{N_{Pr}} ||d_{ij}|| \tag{5.31}$$

where $d_{ij} \in \mathbb{C}^2$ denote the difference between the $j$-th prototype of the previous frequency bin and a prototype in the current bin. The permutation problem is thus solved starting with low frequencies and ending up with high frequencies. The term $\hat{\Pi}$ assigns the prototypes of the current frequency bin to their correspondent HRTF values in the previous bin such that the distance between them is minimum. Figure 5.8 illustrates the assignment of prototypes in adjacent frequency bins.

The problem may arise that, due to little sparseness of the sound sources at a certain frequency, there are less clusters than sound sources present. The number of prototypes $N_{Pr}$ is consequently smaller than $M$. Then, in order to avoid a mismatch of the dimensions of $H(f)$ and $\Pi_i$, $(M - N_{Pr})$ HRTF values of the previous frequency bin are copied to the current bin. these values are supposed to be a good guess for the missing clusters.

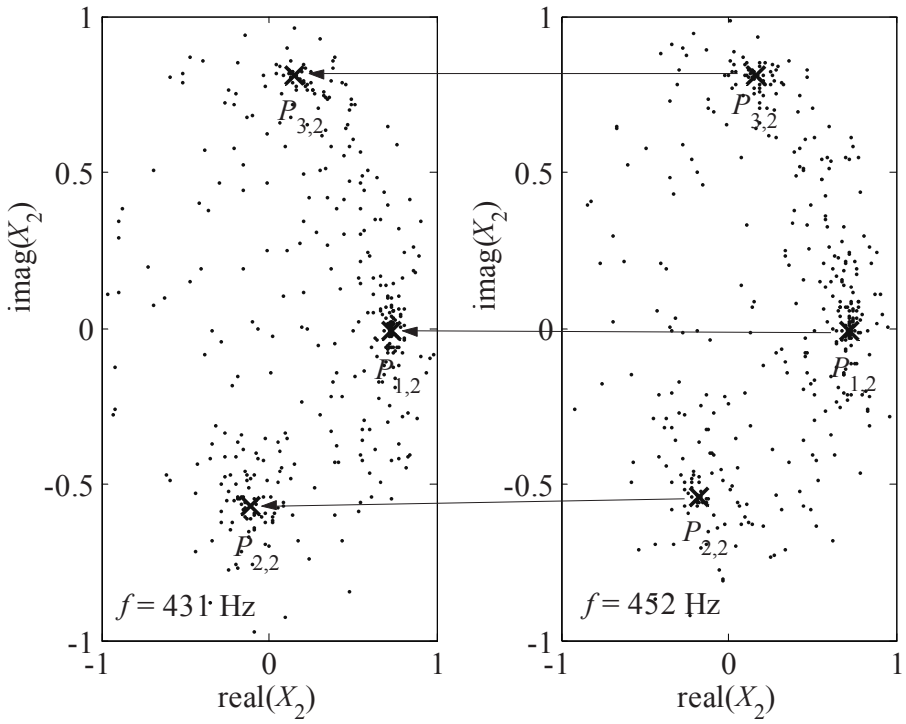


Figure 5.8: In adjacent frequency bins the position of the clusters in the feature space hardly changes. Hence, one searches for the prototype in the previous frequency bin which has minimum distance to a prototype in the current frequency bin. The arrows match two prototypes which belong to the same HRTF.

### 5.3.4 HRTF Database Lookup

In order to determine the azimuth and elevation angles of the concurrent sound sources, we have to find the HRTFs for the left and right ears, inside the KEMAR database, which correspond to our estimated HRTFs. We can calculate, for each sound source, the interaural HRTF $A_{est}(f)$ by dividing our estimated HRTF of the right ear by the HRTF of the left ear. The ILD and IPD are calculated using the expressions $\Delta L_{est}(f) = 20log|A_{est}(f)|$ and $b_{est}(f) = \angle A_{est}(f)$, respectively. The interaural HRTF, of the KEMAR database, denoted by $\hat{A}_{CIPIC}(f)$, that best matches $A_{est}(f)$, is determined as follows:

$$\hat{A}_{CIPIC}(f) \;=\; A_i(f) \min_i \left( \underset{f}{median} |\Delta L_{est}(f) - \Delta L_i(f)| \wedge \underset{f}{median} |b_{est}(f) - b_i(f)| \right) \tag{5.32}$$

Equation (5.32) is evaluated for frequencies within the range 200 Hz to 11 kHz, since in this region binaural cues are very distinct. Having found the correct interaural HRTFs from the database one can determine the azimuth and elevation angles of the sound sources because each HRTF is unique for a certain position in 3D space. Figure 5.9 shows the estimated ILD and IPD that result from the interaural HRTF if a sound source is placed at 30 azimuth and 0 elevation. In order to illustrate the database lookup, the corresponding database ILD and IPD are shown in the right subplots.

## 5.4 Source Separation Process

### 5.4.1 Determined System

Using matrix-vector notation, we can express the windowed and Fourier-transformed ear signals in (5.17) and (5.18) for the case of two concurrent sound sources in a compact form,

$$\begin{bmatrix} X_1(f,\tau) \\ X_2(f,\tau) \end{bmatrix} = \begin{bmatrix} H_{11}(f) & H_{21}(f) \\ H_{21}(f) & H_{22}(f) \end{bmatrix} \cdot \begin{bmatrix} S_1(f,\tau) \\ S_2(f,\tau) \end{bmatrix} \tag{5.33}$$

Obviously, this system is mathematically determined, as the number of equations equals the number of unknowns, which are in this case $S_1(f,\tau)$ and $S_2(f,\tau)$. The HRTFs in the matrix are the database HRTFs found by the HRTF database lookup described in the previous section. Consequently, in order to retrieve the source signals in the time-frequency domain, the matrix in (5.34) simply has to be inverted, whereby we assume that it has full rank due to distinct positions of the two sources in 3D space. So, the sound sources are obtained by

$$\begin{bmatrix} S_1(f,\tau) \\ S_2(f,\tau) \end{bmatrix} = \begin{bmatrix} H_{11}(f) & H_{21}(f) \\ H_{21}(f) & H_{22}(f) \end{bmatrix}^{-1} \cdot \begin{bmatrix} X_1(f,\tau) \\ X_2(f,\tau) \end{bmatrix} \tag{5.34}$$

This inversion is done in each frequency bin which yields the Fourier spectrum of a time frame of the separated speech signals. Afterwards, applying inverse Fourier transform and assembling all time frames with the overlap-add method, we get the separated sources in time-domain. Figure 5.10 is a block diagram illustration of the overall concurrent sound localization system.

### 5.4.2 Underdetermined System

Both equations (5.17) and (5.18) can be further expressed in compact matrix-vector notation as follows:

$$X = \begin{bmatrix} h_1 & h_2 & \dots & h_m \end{bmatrix} S = HS \tag{5.35}$$

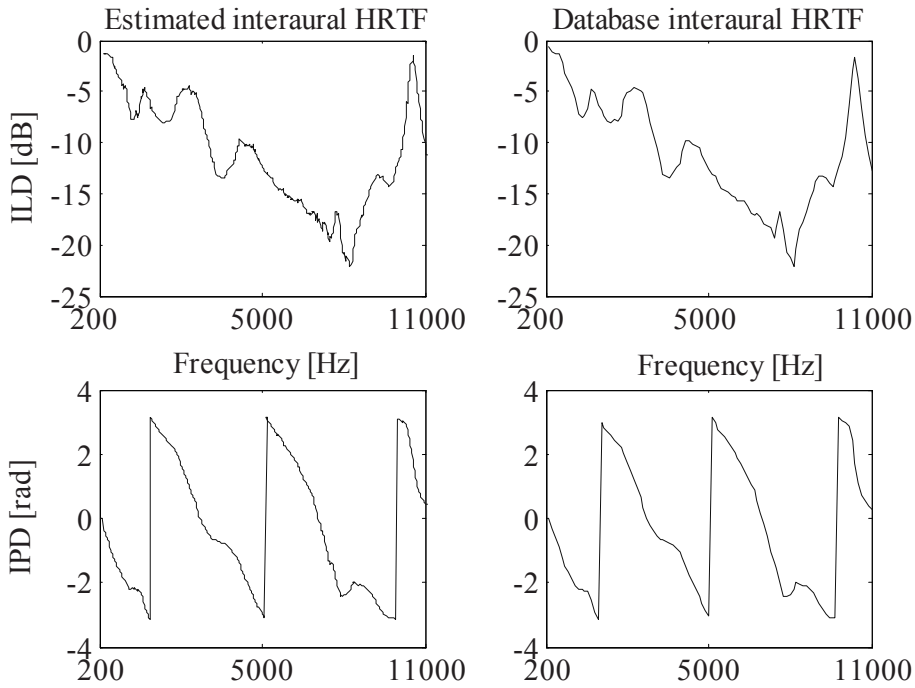


Figure 5.9: Estimated interaural HRTF (left) and corresponding interaural HRTF from database (right). In this case the sound source is placed at 30 azimuth and 0 elevation. In each frequency bin the absolute difference between the estimated ILD/IPD and the database ILD/IPD is calculated. The HRTF from the database which yields minimum difference in ILD and IPD is assumed to be the one that actually filtered the source signal.

where $X = \begin{bmatrix} X_1(f,\tau) & X_2(f,\tau) \end{bmatrix}^T$, $h_j = \begin{bmatrix} H_{j1}(f) & H_{j2}(f) \end{bmatrix}^T$ and the sources matrix $S = \begin{bmatrix} S_1(f,\tau) & S_2(f,\tau) \end{bmatrix}^T$ with $X, h_j \in \mathbb{C}$ and $S \in \mathbb{C}^M$ . If this system of equations was determined in the case that only two sound sources are present, matrix $H$ including the HRTFs could simply be inverted in each frequency bin. However, with three or more sound sources, a more sophisticated method is required since the system of equations is underdetermined and the matrix $H$ is not invertible. The sources can only be retrieved if some more properties of the sound sources are known. With concurrent speech signals, we can assume that their spectral components have statistically independent phases and have amplitudes that are Laplacian distributed. If we follow a Bayesian approach which maximizes the a-posteriori $P(S, H/X)$, we get in the noise-free case and with known HRTFs the cost function [138]

$$\max_{s} P(S) \quad s.t. \;\; X = HS \tag{5.36}$$

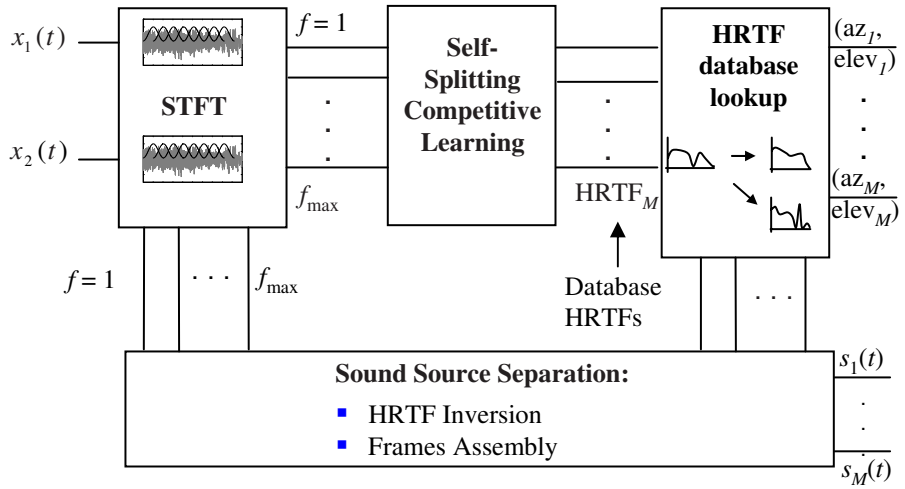


Figure 5.10: After STFT of the ear-input signals, the self-splitting competitive learning algorithm finds the prototypes that represent the HRTFs in each frequency bin. By looking for the HRTFs that match best the estimated ones, the azimuth and elevation positions of the $M$ sources are determined. With the aid of these database HRTFs the sound sources are separated.

With Laplacian distributed components $S_i$, this cost function yields

$$\min_s \sum_i |S_i|, \quad i = 1, \ldots, M \quad s.t. \quad X = HS \tag{5.37}$$

for each time instance $\tau$. In the case, using the L1-norm minimization of real-valued problems, we obtain the sound sources by shortest path decomposition as proposed in [22]. In our particular case with two microphones, the shortest path from the origin to the data point $X$ is obtained by choosing the two HRTF vectors $h_1$ and $h_2$ whose directions are the closest from below and from above to the direction $\theta$. In the example of Figure 5.11 these HRTF vectors are illustrated. The solution $S_{path}$ is then obtained by:

$$S_{path} = \begin{bmatrix} \begin{bmatrix} h_1 & h_2 \end{bmatrix}^{-1} X \\ 0 \end{bmatrix} \tag{5.38}$$

However, the source signals and the HRTFs in frequency domain are in general not real-valued, so that the shortest path decomposition does not necessarily yield the minimum L1-norm. Hence, we add a vector $\hat{S} \in \mathbb{C}^M$ to our solution $S_{spath}$ that is element of the nullspace of $H$ and thus fulfills:

$$X = HS_{spath} + H\hat{S} = H(S_{spath} + \hat{S}) \tag{5.39}$$

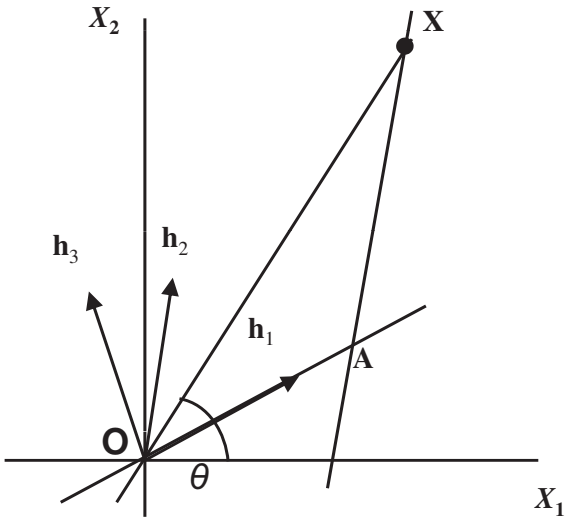


Figure 5.11: The shortest path from the origin **O** to the data point **X** is **O-A-X**. Hence, **X** decomposes as **O-A** along direction $h_1$, as **A-X** along direction $h_2$ and zero along direction $h_3$. The vectors $h_1$ and $h_2$ enclose $\theta$ from above and from below.

For the special case that three concurrent sound sources are to be separated, the nullspace $N(H)$ of $\begin{bmatrix} h_1 & h_2 & h_3 \end{bmatrix}$ is expressed by:

$$N(H) = \alpha \begin{bmatrix} \begin{bmatrix} h_1 & h_2 \end{bmatrix}^{-1} h_3 \\ 1 \end{bmatrix} = \alpha a, \quad \alpha \in \mathbb{C} \tag{5.40}$$

with the complex scaling factor $\alpha$ and the base vector $a$ of the nullspace. The vector of the separated sound sources in time-frequency domain is then obtained by:

$$S_{sep} = \min_{\alpha} |S_{spath} + \alpha a| \tag{5.41}$$

where $|.|_1$ denotes the L1-norm of a vector. The separation algorithm described in this section is performed in all time frames in each single frequency bin. After applying inverse Fourier transform to the time frames, they are assembled with the overlap-add method, which yields the sought separated sound sources in time-domain.

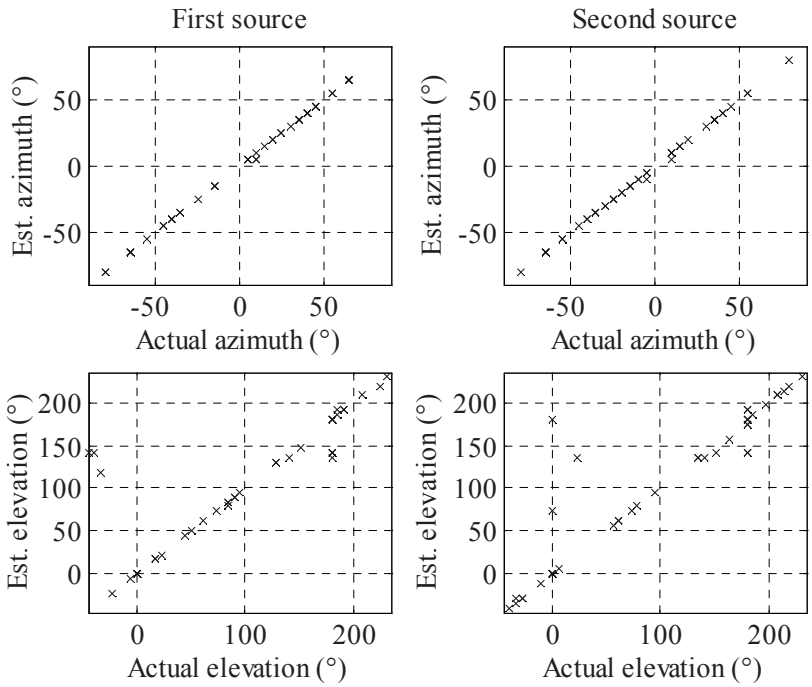


Figure 5.12: Estimated azimuth (top) and elevation (bottom) angles obtained from 50 simulation runs with two male speakers randomly positioned in the horizontal plane and the whole 3D space.

## 5.5 Simulation Results

### 5.5.1 Localization With Adaptive MIMO Systems

In order to assess the performance of localization and separation of two sound sources with an adaptive MIMO system, 50 simulation runs were performed. In the first 20 simulations, the positions of two sound sources were all chosen in the horizontal plane (zero-elevation plane) with random azimuth angles. In the remaining 30 runs, the sources were placed at randomly determined positions in the whole 3D space. The simulation results are depicted in Figure 5.12. The top two subplots show the azimuth values estimated by the algorithm versus the actual azimuth angles of the sources. At the bottom, the corresponding elevation angles are shown. The two sound sources were always positioned in two different hemispheres surrounding KEMAR, i.e. one sound source was on the left and the other one on the right-hand side. If the two sound sources had been placed in the same hemisphere, the algorithm would not have been able to separate, and thus localize them neatly since only causal demixing filters can be trained. In order to localize sources which are in the same hemisphere, noncausal demixing impulse responses would be necessary.

The parameters of the algorithm were chosen as follows. The length of the impulse responses of the demixing MIMO system $L$ was 512 samples. Each of the offline blocks contained 1024 samples, which corresponds to a time frame of 23 milliseconds at a sampling rate of 44.1 kHz. The maximum number of iterations $j_{max}$ in the offline mode was 10 and during one iteration 8 offline blocks were simultaneously processed for decorrelation of the output signals of the MIMO system. For proper adaptation of the demixing filters, 150 online blocks were read in, which is equal to a length of 17 seconds of the microphone signals. Moreover, samples of two male speakers were used as sound sources to evaluate the algorithm. In 74% of all simulation runs, both sound sources were correctly detected with a tolerance of 5° in azimuth and 10° in elevation [79]. As Figure 5.12 shows, the algorithm estimated azimuth values at 180° when the source actually was at 0°, a phenomenon which is well-known as front/back confusion by psychophysical hearing experiments with humans. In nature, mammals tackle this problem by head movements. The algorithm determined the elevation angles quite reliably; only near the horizontal plane (azimuth 0° and 180°), there is a slight deviation from the actual values in some cases. However, humans also exhibit localization difficulties in these regions [19].

### 5.5.2 Localization by Clustering in Time-Frequency Domain

#### Two Concurrent Sound Sources

We tested our new sound localization algorithm by performing 100 simulation runs with two concurrent sound sources located in free space. In 40 simulations, we positioned the sources in the horizontal plane (zero-elevation plane). In half of all the tests, both sounds were situated near each other in the same hemisphere around the KEMAR head. The concurrent sound sources were speech signals of two male speakers sampled at a rate of 44.1 kHz and 16 bit. For binaural synthesis, these mono signals were convolved with the different HRTFs of the KEMAR database, simulating thus different locations in space.

The ear input signals were windowed with a Hamming window of 1024 samples and an overlap of 50% used for properly calculating the STFTs. For clustering, 400 time-frames of the ear input signals were acquired by the algorithm. This resulted in a signal length of approximately 4.7 seconds. Figure 5.13 shows the results of the 100 simulation runs. The left subplot depicts the estimated azimuth (above) and elevation (below) angles versus the actual ones for the first speaker (left) and the second speaker (right). Notably, the observed localization rate was 100%. For all the simulation runs, both concurrent sound sources were located exactly at their target azimuth and elevation angles. The algorithm showed the same 100% localization

performance in the case of both sound sources located close to each other in the same hemisphere around the KEMAR head.

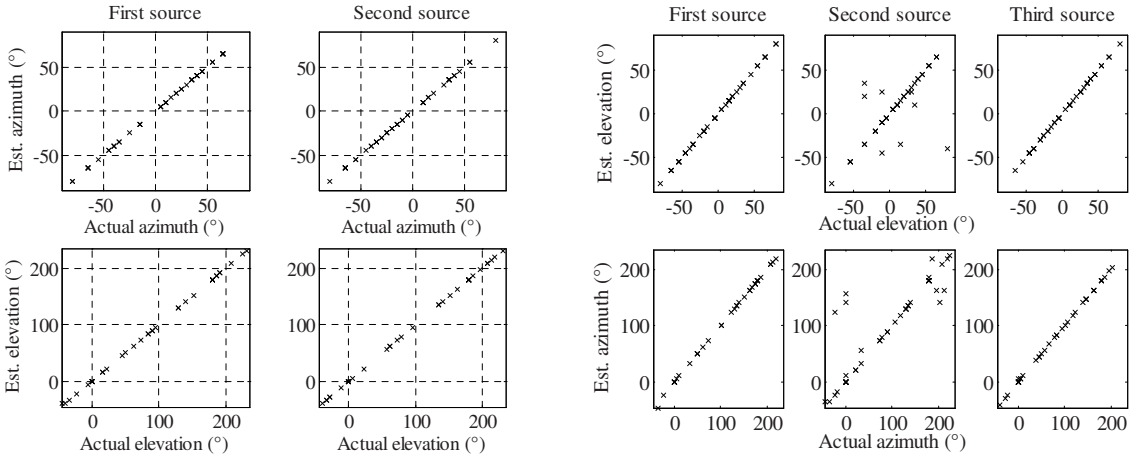


Figure 5.13: **Left:** Estimated azimuth (top) and elevation (bottom) angles obtained from 100 simulation runs with randomly chosen two concurrent speaker positions in the whole 3D space. **Right:** Estimated azimuth (top) and elevation (bottom) angles for three concurrent speakers.

In further 50 simulation runs, a stationary noise source (computer fan) was introduced. This noise source was constantly located at 0 azimuth and 0 elevation. The mean SNR, i.e. the ratio of the mean power of the speech signals to the noise power was chosen to be 20 dB. Under these conditions, the localization percentage fell to 90%. For 25 dB SNR, the algorithm is quite robust to the stationary noise since the localization accuracy rises to 97% [81].

### Three Concurrent Sound Sources

A total of 50 runs with three concurrent, randomly positioned sound sources were simulated. In 20 simulations, we positioned the sources in the horizontal plane (zero-elevation plane). The azimuth angles were always chosen in such a way that two of the sound sources were virtually placed in the same hemisphere (both to the left or to right of the humanoid), and the third one in the other hemisphere. The concurrent sound sources were speech signals of two male speakers and one female speaker sampled at a rate of 44.1 kHz and 16 bit. The ear input signals were windowed by the same Hamming window as before. For clustering, 400 time-frames of the ear input signals were acquired. Figure 5.13 shows the results of the

50 simulation runs. The right subplots depict the estimated azimuth (top) and elevation (bottom) angles versus the actual ones for the first (left), second (middle) and third (right) speaker. The simulations yielded an overall correct localization rate of 78%, a remarkable result, especially that two of the three concurrent sources are always perfectly located at their target location [80].

Table 5.1: Signal to Interference Ratio (SIR) for determined and underdetermined sound source separation.

| | **Mean** $SIR$ | $SIR_A$ | $SIR_B$ |
|---|---|---|---|
| **Adaptive MIMO (determined)** | 25.2 dB | 19.2 dB | 33.5 dB |
| **Clustering with 2 sources** | 37.2 dB | 36.5 dB | 37.9 dB |
| **Clustering with 3 sources** | 21 dB | 13.1 dB | 31.1 dB |

### 5.5.3 Separation Performance

The quality of separation was also assessed in several simulations. An appropriate measure for the separation performance is the signal-to-interference ratio (SIR) of the output signals of the algorithm. As proposed in [129], the SIR is calculated as follows:

$$SIR = 10log_{10}\frac{\|s_t\|}{\|e_i\|}, \tag{5.42}$$

where $s_t = \frac{<y_i,s_j>s_j}{\|s_j\|^2}$ and $e_i = \sum_{j' \neq j} \frac{<y_i,s_{j'}>s_{j'}}{\|s_{j'}\|^2}$. Equation (5.42) represents the ratio between the power of the desired separated signal $s_t$, to the power of the interfering signal $e_i$, in the $j$-th output channel. The term $< \cdot, \cdot >$ denotes the inner product of two signals. The variables $y_j$ and $s_j$ denote the time signal of the $i$-th output signal of the separation system and the $j$-th sound source, respectively. The index $j'$ denotes the other source signals which contribute to the interference signal $e_i$.

The median SIR value obtained after running 100 simulations is 37.2 dB. In statistical terms, 50 % of the SIR values lay between 36.5 dB and 37.9 dB. Compared to other blind source separation methods, e.g. [57], trying to solve the same determined problem, our separation algorithm yields an average SIR that is more than 10 dB higher.

Finally, we investigated the performance of source separation described above in the noise-free case. Towards this end, we used the above-defined SIR ratio. Calculating the median of the SIR values for all 50 simulations, we observed an SIR value of 21 dB. In statistical terms, 50% of the SIR values lay in the acceptable

interval of 13.1 dB to 31.2 dB. These values are notably higher than SIR values provided by other underdetermined separation algorithms, trying to solve the problem of more sources than observations under the same conditions [25]. The sound source separation performance bounds for the determined and underdeterined cases are summarized in Table 5.1

The interval between $SIR_A$ and $SIR_B$ encloses 50% of the SIR values around the median value. In case of three sound sources, the obtained values are higher than SIR values provided by other underdetermined separation algorithms, trying to solve the problem of more sources than observations, e.g. [20], since the measured database HRTFs and not the estimated HRTFs are used for source separation [70].

### 5.5.4 Localization Performance

We have compared our method with the concurrent detection method based on the echo-avoidance model [51] introduced in chapter 1. Sound localization experiments were conducted in an anechoic chamber and a normal room. Three microphones arranged in a triangular form were available for sound recording. Two sound sources were played concurrently. The sources were set in two fixed positions each distanced 1 meter from the wall. The azimuth of the first sound source was 0 degree facing direction one of the three microphones and the second sound source was about 38 degrees to that microphone. The distance from the sound sources to the center of the microphone set was about 2.9 m and the distance between sound sources was about 1.9 m. Sound data were recorded by a multi-channel analog data recorder with a sampling frequency of 9600 Hz.

The azimuth histograms for the anechoic chamber and the normal room were computed. The contribution of each sound/echo was added to the sum azimuth histogram. The histogram peaks are pronounced around 0 and 38 degrees over all time segments in the histograms of the anechoic chamber. These two rows of peaks correspond to the first and the second sound source, respectively. The positions of major peaks are in the regions of [0, 4] and [35, 39] degrees, i.e., the first sound source was localized in 2($\pm$2) degrees and the second sound source in 37($\pm$2) degrees. The maximum absolute error is 4 degrees. The histograms of the normal room, however, show more disorder comparing to the anechoic chamber. The scores are smaller and the size of the peaks is not consistent. After smoothing the histograms by a two-dimension Gaussian function, the time resolution decreased to about 2 s, but the peak positions became more consistent. The positions of major peaks are in the regions of [-2, 2] and [33, 37] degrees, i.e., the first sound source was localized in 0($\pm$2) degrees and the second sound source was localized in 35($\pm$2) degrees. The maximum absolute errors is 5 degrees. All the results are similar to the results

obtained in the anechoic chamber except that 2 seconds are now needed to obtain an accurate localization, compared with 0.5 second for the anechoic room. Using our SSCL-based localization method in the anechoic room, the first sound source located at 0 degree was detected in the region [0, 5] degrees. The second source located at 38 degrees was detected in the region [35, 40] degrees. The maximum absolute error is 5 degrees and the overall processing time is 4.7s. In the normal room, the first sound source located at 0 degree was detected in the region [-5, 0] degrees, and the second source located at 38 degrees was detected in the region [30, 35] degrees. The maximum absolute error is 5 degrees for and overall processing time of 11.5s. It should be noted that, on the expense of increased computational power, the 5-degree absolute error could be reduced to 1 degree if we use a HRTF database sampled every one degree in the azimuthal plane. In addition, the SSCL-based method is not limited to azimuthal localization and uses only two microphones.

Furthermore, we compare our method with the beamforming technique [102] introduced in chapter 1. This method utilizes a microphone array of 8 sensors and deploys time delay of arrival estimation, and multiple Kalman filters for concurrent sound source tracking in azimuth. A steered beamformer was used for the sound source localization. The basic idea of this method is to direct a beamformer in all possible directions and look for maximal output. The beamformer searches a spherical space around the microphone array which is divided into 5,120 triangle grids with 2,562 vertices. The beamformer energy is computed for each vertex by incremental refinements from a large triangle to smaller ones. The direction of a sound source is estimated as that of the region with the maximal energy. This method localized sound source accurately for stationary and moving sound sources. Although the method provided directional information at each time frame, a temporal grouping of the same sound source was not attained. Therefore, it was difficult to track multiple moving sound sources. This ambiguity was clarified by feeding the temporal information to a multiple Kalman filter with different history lengths. Multiple Kalman filters with different history length predict next states in parallel, and provides a set of estimates. The current estimate is obtained by the filter which predicted the state with the minimal error in the previous frame. The continuity of localization for each speaker is forced by using acoustic features of separated speech signals. For this purpose, the power spectrum was computed in each frame, in order to reduce the ambiguities in tracking moving speakers. The observed value whose power spectrum is similar to the past one of a speaker is selected as the observed value of the speaker. The power spectrum of a separated sound is calculated by using a delay-and-sum beamformer, which uses a localization angle as a clue. If the separated speech of each moving speaker is available, the fundamental frequency of the speech is used for the accurate selection of the observed value.

The algorithm processes loudspeaker signals using 8 channel sound recorded

with a sampling rate of 48kHz. The accuracy of the multiple Kalman filter method, for two concurrent sound sources placed at 0 and 30 degrees in front of the robot, yielded a mean square angular error of 26, compared to a mean square error of 20 for our SSCL-based technique. For three concurrent sound sources located at 0, 80 and 110 degrees, the multiple Kalman filter method undergoes a mean square angular error of 28, compared to a mean square error of 25 for the SSCL-based technique. The increased accuracy of the SSCL-based technique comes at the expense of a 4.7s processing time compared to 2.4s for the multiple filter method. While the multiple Kalman filter method covers only azimuthal localization, our SSCL-based method covers the whole 3D space surrounding the robot head, and deploys two instead of 8 microphones.

Summing up, we have used only two microphones in combination with a generic HRTF database to localize and separate two concurrent moving sources in a 3D space. The simplicity of the proposed sound source separation and localization algorithm suggests a cost-effective implementation for robot platforms. While the detection of sound in elevation was similar to the human ability, our results demonstrated very precise localization of the sound sources in the azimuth angles. Since the HRTF dataset used was measured on the horizontal plane from 0° to 360° with 5° increment and on the vertical plane from -40° to 90° with 10° increment, we can localize the sound source with an accuracy of about 5° to 10°. Obviously, an HRTF dataset with smaller increments increases the resolution of estimation. In the following chapter, the HRTF interpolation method presented in chapter 3 will be used to enhance the accuracy of a localizer operating in a highly-reverberant environment.

# Chapter 6

# Sound Localization in Highly Reverberant Environments

In this chapter, we introduce a robust sound localization algorithm, which uses Bayesian information fusion to increase the localization resolution in a three-dimensional reverberant environment. The main focus is the detection of sound events under severe acoustic conditions, i.e. high reverberation and background noise. The location of the sound source obtained from a number of observation sensors is fused using a properly tuned Bayesian network so that an accurate three-dimensional direction of arrival estimation, in terms of both azimuth and elevation, is guaranteed.

## 6.1 New Hardware Setup

The new sound localization setup takes two sound signals as input: 1) the spatial sound signal measured inside the ear canal of KEMAR humanoid's artificial ear, and 2) the sound signal measured outside the artificial ear, placed 5 cm away from the inner microphone. This hardware configuration is illustrated in Figure 6.2.

After data acquisition, both inner and outer signals are divided, in the spectral domain, in an attempt to exclude the incoming sound signal and, thus, isolate the effect of the pinna, head, and torso. Consequently, the appropriate HRTF which has shaped the incoming sound signal is extracted. Using simple correlation, the extracted HRTF is then compared with a database of HRTFs, [6], and the maximum correlation coefficient is taken to be corresponding to the 3D sound source location. The HRTFs were measured every $5^\circ$ in elevation and azimuth. An accurate, recently

proposed HRTF interpolation method [66], see chapter 3, is then used to obtain a high-spatial-resolution HRTF database with one HRTF every 1° azimuth, spanning an elevation range from -20° to 60°. Each of the 28800 HRTFs 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.

## 6.2 Monaural System

Our proposed monaural sound localization system receives two input signals collected at two small microphones, one inserted inside and one placed outside the artificial humanoid ear. The left and right blocks of Figure 6.2 illustrate the monaural localization at both ears.

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:

$$\mathrm{S}_{in\_L}(f) = \mathrm{S}^{c}_{out}(f) \cdot \mathrm{H}_{ss} + \sum_{i=1}^{N} E_{in\_i}(f).\mathrm{H}_i + n \tag{6.1}$$

where $\mathrm{S}_{in\_L}(f)$ is the signal received at the microphone inside the ear, $S^{c}_{out}(f)$ is the clean sound signal arriving at the ear canal, $\mathrm{H}_{ss}$ is the correct frequency shaping response corresponding to the location of the source, and $E_{in\_i}(f)$ is the $i^{th}$ echo inside the ear arriving from some position in 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 $\mathrm{H}_i$ 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:

$$\mathrm{S}_{out\_L}(f) = \mathrm{S}^{c}_{out\_L}(f) + \sum_{i=1}^{N} E_{out\_i}(f) + n_s \tag{6.2}$$

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

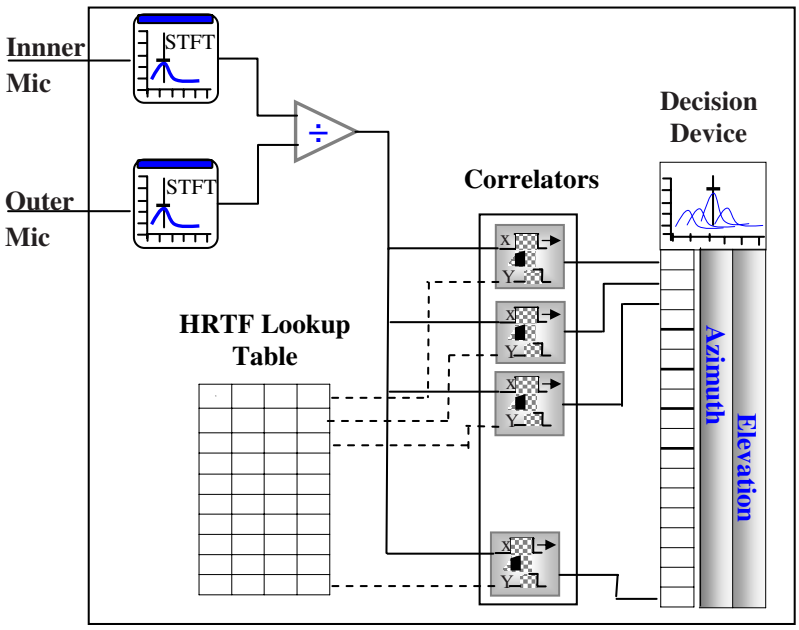


Figure 6.1: Proposed localization mechanism: After data acquisition, both inner and outer microphone signals are divided.

Dividing both Equations 6.1 and 6.2, and assuming that the echo signals received are attenuated considerably, the term $H_{ss}$ dominates the division operation result. Theoretically speaking, in a noise-free anechoic environment, the division operation would result only in $H_{ss}$.

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 compared at the $i^{th}$ correlator with the $i^{th}$ HRTF available 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. This localization mechanism is illustrated in Figure 6.1. The same procedure is repeated for the right ear [84].

## 6.3 Combined System

In the binaural localization case, we use the cross convolution system introduced in chapter 4. In this context, the received signals at the microphones inside the ear

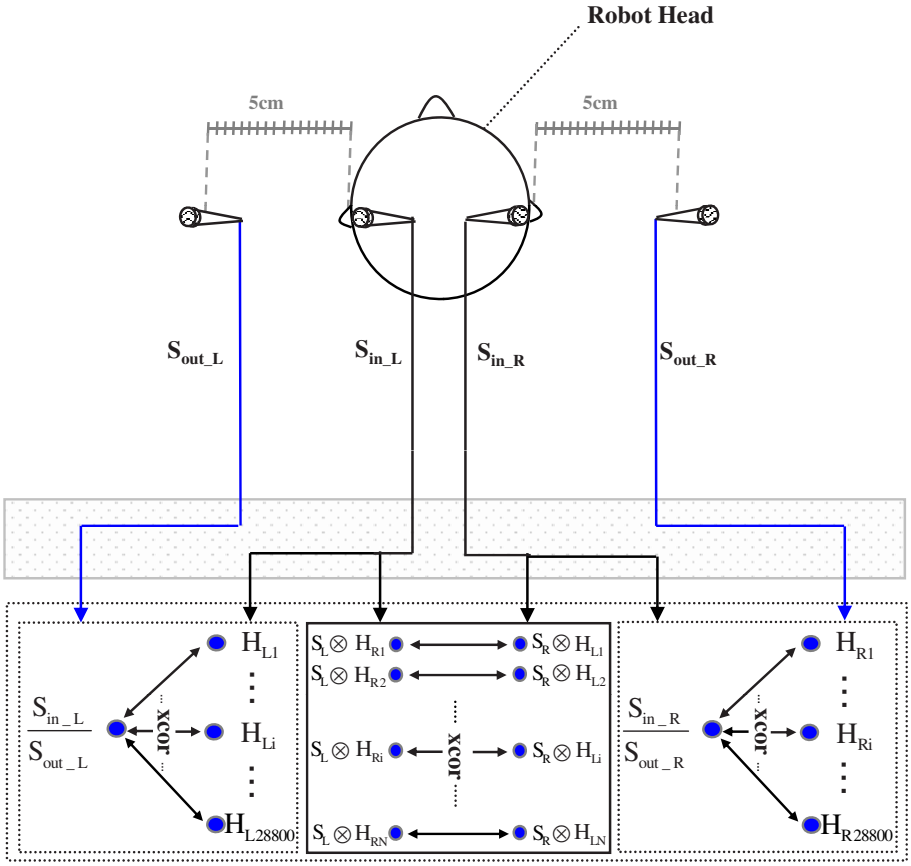


Figure 6.2: Block diagram of the overall localization system.

canals are modeled as the original sound source signal convolved with appropriate HRTFs. The left and right microphone signals are multiplied by the right and left HRTFs, respectively. The sound source location is estimated by finding the maximum correlation coefficient between incoming and saved HRTFs.

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 case 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 [83]. This is however a lossy method to combine monaural and binaural estimations since we are discarding a useful part which could otherwise contribute to train the system. The localization method could be made intelligent by learning a-priori information from training data, and thus minimizing the uncertainty of the online localization. For

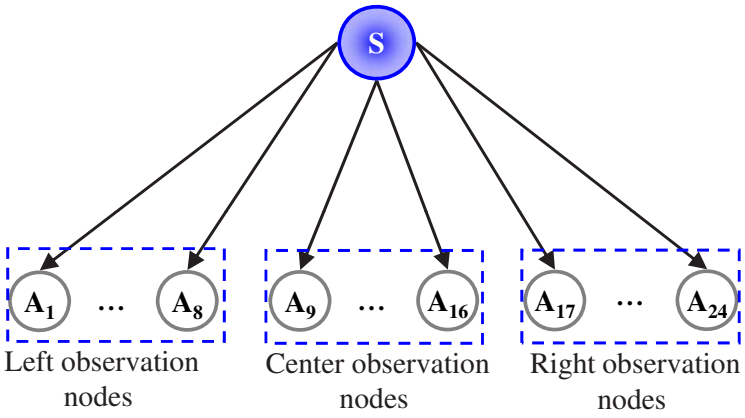


Figure 6.3: Proposed Bayesian network for the monaural and binaural information fusion.

this purpose, a Bayesian network is deployed.

## 6.4 Bayesian Information Fusion

### 6.4.1 Feature Vectors Extraction

A Bayesian network is employed to detect the 3D direction of the instantaneous sound event. The Bayesian network is a way of modeling a joint probability distribution of multiple random variables and is considered to be a powerful tool for information fusion [55]. Figure 6.3 shows the topology of the Bayesian network we have adopted in our work. The network has $N = 24$ nodes, $\{A_1, ..., A_N\}$, divided into three sectors: left, center, and right. The left eight nodes, $\{A_1, ..., A_8\}$, correspond to the left monaural system, and represent the first eight estimated locations having the first maximum correlation coefficients. The center nodes $\{A_9, ..., A_{16}\}$ correspond to the binaural system. Similar to the left nodes, they represent eight estimated locations which have the maximum correlation coefficients. The remaining nodes $\{A_{17}, ..., A_{24}\}$ correspond to the right monaural system.

The left, center, and right sectors are then compared to check which of their entries match. For every matching node, i.e. if one node of the left monaural set of nodes is found in the binaural node set, this information is converted to a state of 0/1 ("1" corresponds to a match being detected). Proceeding this way, the feature vector $a(t) = A_1(t), ..., ..., A_N(t)$ is formed. Figure 6.3 shows the Bayesian network used for fusing the left, center and right observation node sectors depicted above.

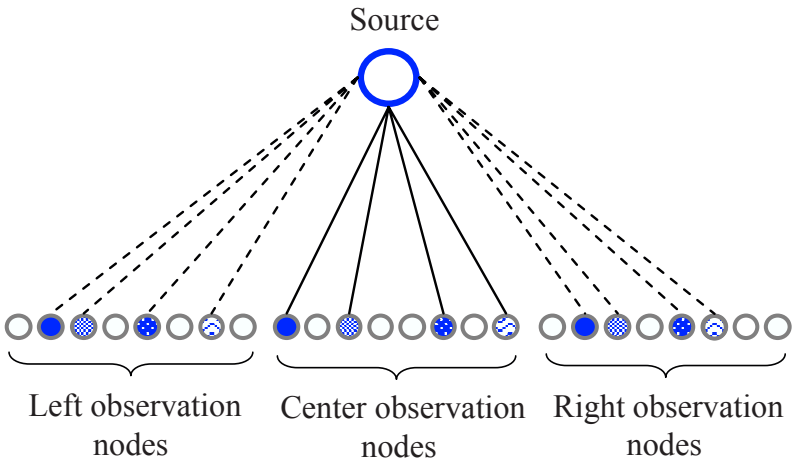


Figure 6.4: A state of the Bayesian network: the connections between the nodes simply emphasize that the corresponding observation nodes are in the "1" state.

### 6.4.2 Decision Making

As previously mentioned, the input nodes have the states of $\{0, 1\}$ according to the occurrence of the corresponding node in more than one node sector. On the other hand, the output node $S$ has the following $N_s + 1 = 28800 + 1$ states: $S = \{S_1, ..., S_{Ns}, NoEvent\}$. The state $S_1, ..., S_{N_s}$ corresponds to the 3D sound position (azimuth, elevation). When $S = (10, -20)$ for example, the speaker is located in the direction of +10 azimuth and -20 elevation and is speaking. When $S = NoEvent$, there are no sound events. For estimating $S$ from the audio feature vectors described before, a properly tuned Bayesian network is used. Figure 6.4 shows a state of the Bayesian network used in this work.

We assume that the value of all $A_i$ are conditionally independent when the value of $S$ is given. Hence, the conditional probability distribution $P(S|A_1, ..., A_N)$ can be factored in the product of local conditional probabilities $P(A_i|S)$:

$$P(S|A_1, ..., A_N) = \frac{1}{Z} P(S) \prod_{n=1}^{N} P(A_n|S) \tag{6.3}$$

where $Z = \int_S P(S) \prod_{n=1}^{N} P(A_n|S) dS$.

The conditional probabilities $P(A_i|S)$ can be estimated from training samples. These probabilities are then saved in a so-called Conditional Probability Table (CPT). For the training samples, the value of $S$ is given as a label for each feature vector. In this work, broadband sound signals from a loudspeaker in a reverberant

laboratory room were used as training samples. The location of the loudspeaker was varied between $-40$ and $+60$ every 5°in azimuth and elevation. Each sample was 30 seconds long. There were no significant noise sources in the laboratory room, there was however high background noise such as that from a PC fan. These samples were used as a supervisor for training the CPT.

In the operation phase, the feature vectors are obtained as evidence at every time block. Using the evidence and the CPTs obtained above, the conditional probability (6.3) is calculated and the most probable state of $S$ is obtained, i.e. the most probable sound source location.

## 6.5 Discussion of Results

### 6.5.1 Simulation 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. White Gaussian noise and high reverberations, i.e. echoes 20dB below the signal level, were added to the simulated sound sources. 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. Basically, the correlation should yield a maximum value when the saved HRTF ratio, for the binaural system, corresponds to the location from which the simulated sound source is originating. Similarly, for the monaural system, when the saved HRTF ratio corresponds to the location from which the simulated sound source is coming, the correlation should yield a maximum value. Therefore, we base our localization on the obtained maximum correlation factor. The PCA reduction technique was used to create a truncated model of the original HRTFs.

To simulate the reverberation in our room environment, the image method for room acoustics was used [7]. 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 simulation setup and room dimensions were defined to match the experimental room environment. 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 source's and microphone's positions. After recombining the convolution results, random Gaussian noise was finally added to each microphone signal yielding an SNR level of 20dB. Figure 6.5 shows the sound localization performance for the cross convolution technique, compared to the combined system operating with Bayesian information fusion. The

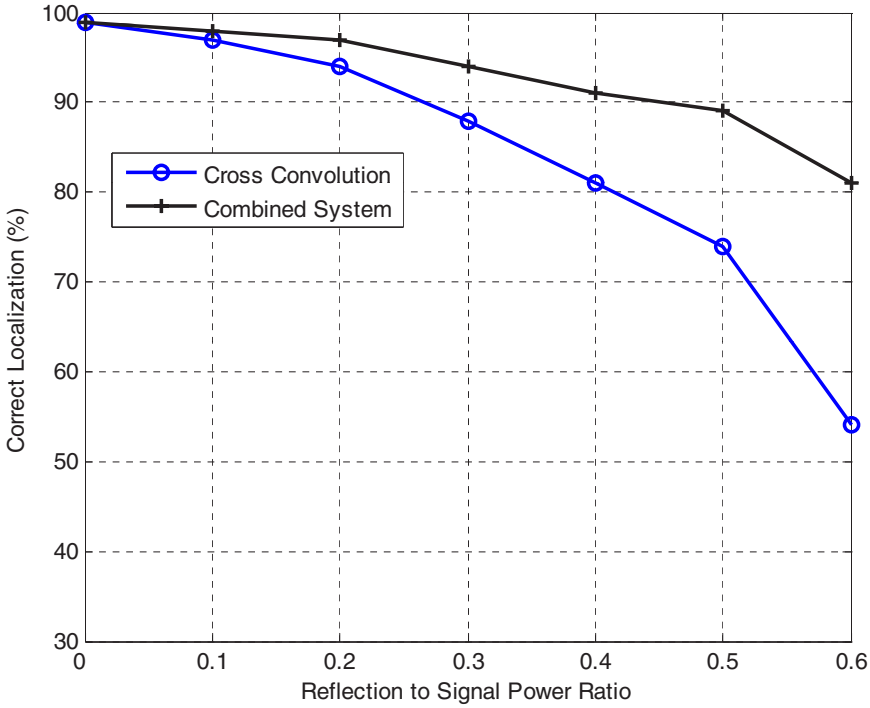


Figure 6.5: Percentage of correct localization for the combined system compared to the cross convolution system. The audio data was simulated with the image method for room acoustics.

HRTFs used in this case are DFE-reduced, i.e. $\mathrm{H}_{128}^{FIR}$. As depicted in Figure 6.5, the combined system is outperforming the stand-alone cross convolution technique especially for high reflection to signal power ratios.

Under high reverberation conditions, i.e. for a reverb time of $RT = 1s$. using the $\mathrm{H}_{m}^{FIR}$ PCA-reduced dataset, the combined system, without Bayesian fusion, yielded a percentage of correct localization between 22% to 81% with the HRIR being within 10 to 45 samples, i.e. $10 \leq m \leq 45$. With Bayesian fusion the localization falls between 31% and 90% for $10 \leq m \leq 45$. For a full-length HRIR of order 512, the percentage of correct localization reached 97% under the same reverberation conditions without Bayesian fusion, and 99% with fusion.

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, the falsely localized angles are from their target location, can be seen in Figure 6.6. The dashed lines and the rigid lines correspond to the simulation and experimental results, respectively. The Figure shows the performance of the localization system with (circles) and without (squares) Bayesian fusion.

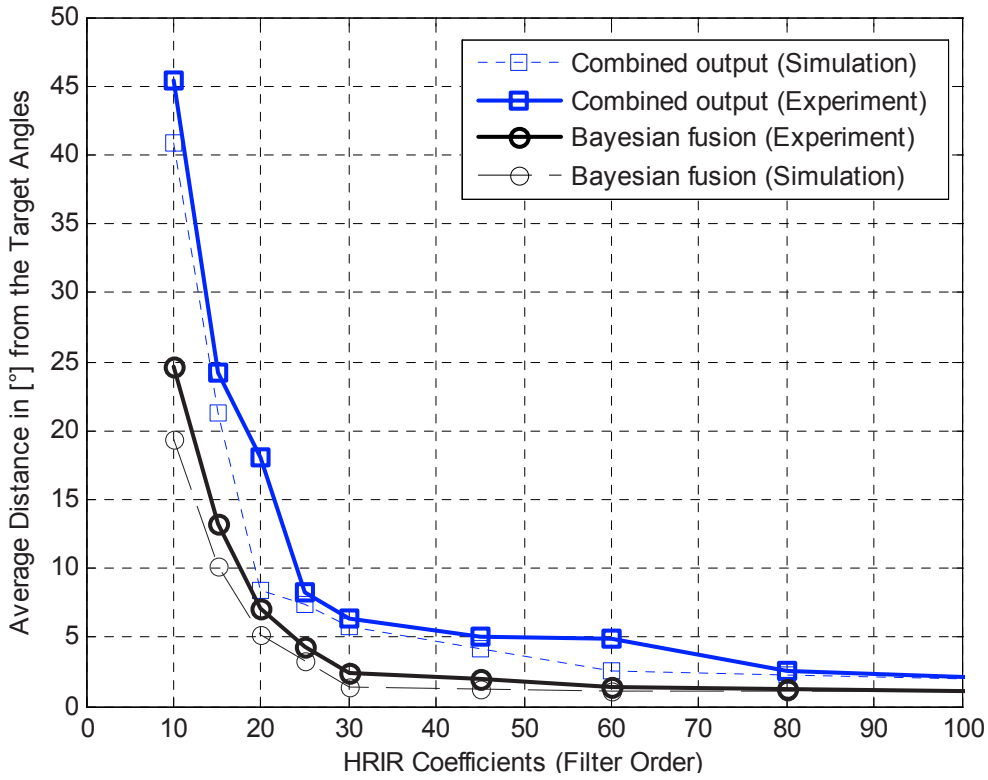


Figure 6.6: Average distance of the falsely localized angle locations to their target positions, for every HRIR filter order.

Intuitively, with more intelligence encapsulated within the HRIRs, the localization accuracy increases. Hence, with more HRIR samples, the average distance to the target sound source location decreases. The combined system, without fusion, reports a worst angular error of 40.83° with a HRIR order of 10, and a best angular error of 2.0° with a HRTF order of 100. With Bayesian fusion, the worst angular error reached 19.33° with a HRIR order of 10, and the best error 1.0° with an HRTF order of 100.

Finally, we have used the million instructions per second (MIPS) measure for the verification of real-time implementation in terms of CPU usage. Table 6.1 shows the different processes underlying the combined system with Bayesian fusion. The algorithm processes 350 msecs of input data sampled at 44.1kHz. The HRTFs used are PCA-reduced using 45 eigenvectors. The kernel of the suggested combined algorithm requires a total of 410,758,808 MIPS. For a 32-bit Pentium IV, 1.9 GHz platform, this corresponds to 220 ms of processing delay. Hence, a real-time sound localization is ensured.

Table 6.1: The number of instructions required for processing 350 msecs of audio input using the combined system.

| **Process** | **MIPS** | **Percentage** |
|---|---|---|
| Audio Buffer Handling | 361,550 | 0.20% |
| Left and right Divisions | 86,110,053 | 20.96% |
| Correlation | 324,287,205 | 78.84% |
| Total | 410,758,808 | 100% |

### 6.5.2 Experimental Results

In our household experimental setup, 100 different binaural recordings were obtained using a broadband sound signal. The speakers were placed 2 meters away at different angle locations around the KEMAR head equipped with two small artificial ears in a highly-reverberant room. Our laboratory setup is illustrated in Figure 6.7. The speaker was held at a constant distance of 1.3 meters from the head. The recording environment was a laboratory room measuring where the walls, ceiling, and Floor are made of unpainted concrete. One wall has a $5m \times 2m$ glass window and is facing the dummy head. The dummy head and torso are placed on a rotating table in the middle of the room. The dummy head artificial ears and microphones are held at a constant height of 1.5 meters from the floor. The room contains objects like tables, chairs, and computer screens.

The level of reverberation in the room was experimentally measured by means of a loudspeaker emitting a high level white noise signal. Measuring the 60dB decay period of the sound pressure level after the source signal is switched off, for a number of speaker and microphone positions, provided the frequency-averaged reverberation time $RT = 1s$.

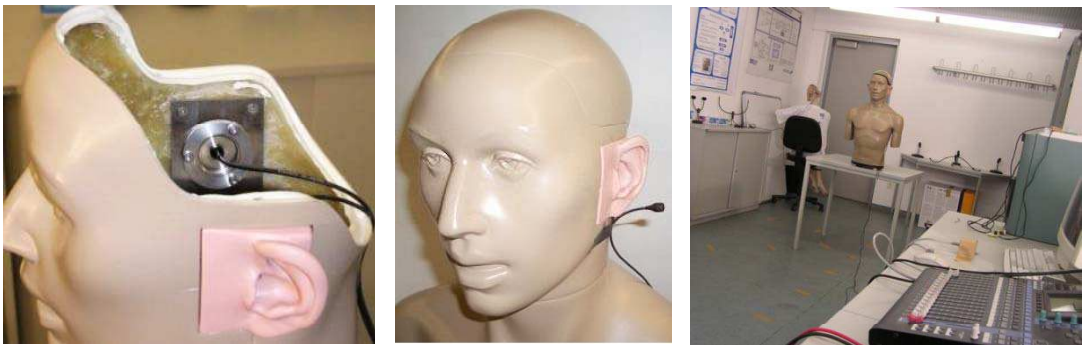

Figure 6.7: The laboratory hardware setup.

To keep a fair comparison with the simulation setup, each of the recordings was 350 $m$secs long, and the noise 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 placed 5 cm away from the inside ones. The recorded sound signals, also containing external and electronic noise, were used as inputs to the system. A HRIR database reduced using the PCA method, $\mathrm{H}_m^{FIR}$, was available for the test.

The combined system, without Bayesian fusion, yielded a percentage of correct localization 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. With Bayesian fusion the localization falls between 19% and 85% for $10 \leq m \leq 45$. For a full-length HRIR of order 512, the percentage of correct localization reached 89% under the same reverberation conditions.

Similar to the simulation results, for high order HRIRs, the falsely localized angles fall in the vicinity of the target sound source. Figure 6.6 illustrates the average distance to the target angles. The combined system yielded a worst angular error of 45.33° with a HRIR order of 10, and a best angular of 1.6° with a HRIR order of 128. With Bayesian fusion, the worst angular error reached 24.57° with a HRIR order of 10, and the best error 1.2° with HRTF order of 100.

It is worth mentioning that common robotic sound localization methods which use only 4 microphones fail to localize sound accurately in three dimensions without becoming impractically complex, or without using computer vision to augment the acoustic modality [56]. We have thus compared our experimental results to the method in [128], (also see chapter 1). This method uses 8 microphones and applies the simple TDOA algorithm to localize sound sources in three dimensions. Like in [128], the sounds we have used have a large bandwidth, e.g. fingers snapping and percussive noises.

The system was tested with sound sources placed at different locations in the environment. In each case, the distance and elevation are fixed and recordings are taken for different horizontal angles. The mean angular error for every arrangement is computed. Table 6.2 shows the performance of our system as compared to the system in [128]. Using only 4 microphones, our system performed more accurately when localizing the sound sources placed at the same distance, azimuth and elevation angles as in [128]. It should be noted that part of this error, mainly at short distances, is due to the difficulty of precisely placing the source and due to the fact that the speaker employed is not a point source. Other sources of error come from reverberation on the floor especially for those locations where the source is high. The angular error is almost the same for sources located in the horizontal plane and

Table 6.2: Localization mean angular error comparison.

| **Distance** | **Elevation** | **Mean Error as in [128]** | **Mean Error (Combined System)** | **Mean Error (Bayesian System)** |
|---|---|---|---|---|
| 3 m | -7° | 1.7° | 1.6° | 1.1° |
| 3 m | 8° | 3° | 1.7° | 1.0° |
| 1.5 m | -13° | 3.1° | 1.9 ° | 1.2 ° |
| 0.9 m | 24° | 3.3° | 2.4° | 1.7° |

varies only slightly with the elevation, due to the interference from floor and wall reflections. This is for example an advantage over the system in [107] where the error is high when the source is located on the sides. Moreover, for the case where multiple sound sources are concurrently active in the humanoid's environment, the SSCL clustering algorithm proposed in chapter 5 could be used for sound source separation and localization [76].

To conclude, we have presented 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 monaural and binaural HRTF cues [73, 68]. 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.

# Chapter 7

# Conclusion

It is a difficult challenge to use only one pair of microphones on a robot to mimic the hearing capabilities of humans. This task is made even more challenging by the fact that the listening environment is dynamic: sound sources appear, disappear, move and interfere with each other. Most of the proposed localization models today are based on using more than two microphones to detect and track sound in a real environment. Mathematical models of sound wave propagation 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 methods are designed to mimic the human biological sound localization mechanism by building models of the outer, middle and inner ear, using knowledge of how acoustic events are transduced and transformed by biological auditory systems. The difficulty with this approach is that neurophysiologists do not completely understand how living organisms localize sounds. For instance, It remains unclear whether the ITD and ILD cues are combined, in the central nervous system, before or after they are spatially mapped. Moreover, 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.

We have proposed a unifying framework for three-dimensional sound localization methods to be deployed on a humanoid robot operating in a general telepresence environment. Motivated by the important role of the human pinnae to focus and amplify sound, and knowing that the HRTFs can also be interpreted as the di-

rectivity characteristics of the two pinnae, only two microphones in combination with a generic HRTF database were required to localize sound sources in a three dimensional space. Common binaural sound localization methods using only two microphones fail to localize sound accurately in three dimensions without becoming impractically complex, or without using computer vision to augment the acoustic modality.

For faster localization performance, the HRTFs are reduced using three different model truncation techniques, namely Diffuse-Field Equalization, Balanced Model Truncation, and Principle Component Analysis. Furthermore, for a robust and more accurate localization mechanism, which demonstrates precise azimuth and elevation estimation, we have introduced a novel state-space solution to the HRTF inversion and interpolation problems. Beside its application in our sound localization system, the stable inversion of transfer functions is of valuable importance for sound synthesis and channel equalization, not only to compensate from deficiencies of the transduction chain (amplifiers, loudspeakers, headphones), but also to reproduce a spatially coherent sound field. On the other hand, the HRTF interpolation technique we have introduced for sound localization purposes can also be used in the immense field of binaural sound synthesis for high-fidelity reconstruction of HRTFs especially in cases where fast and immersive synthesis of moving sound sources is needed.

The initially proposed sound localization method is based on dividing the ear signals with the left and right HRTFs and subsequently taking the maximum correlation coefficient as a pointer to the source location. This method is enhanced using proper state-space HRTF inversion. Nevertheless, a new algorithm called cross convolution was developed to further decrease the computational requirements of the initial method. In comparison to the previous methods, the cross convolution is able to achieve remarkable reduction in the processing requirements while increasing the accuracy of the sound localization. Furthermore, with the help of a simple properly tuned Kalman filter, a ROI was introduced to account for fast moving sound sources. Simulation and experimental results showed a real-time tracking performance and a higher noise-tolerance capacity. The efficiency of the new algorithm suggests a cost-effective implementation for robot platforms and allows fast localization of moving sound sources.

Using the presented methods, we have addressed the challenging task of binaural concurrent sound source localization and separation in reverberant environments. Relying on the concept of binaural hearing, where the human auditory 3D percepts are predominantly formed on the basis of the sound-pressure signals at the two eardrums, our robotic localization system uses only two microphones. We presented a new algorithm for binaural localization of concurrent sound sources in both azimuth and elevation. By exploiting the ILD and IPD binaural cues that

are encapsulated within the HRTFs, binaural 3D concurrent sound localization was made possible using only two microphones placed inside the artificial ears of the KEMAR head. Compared to existing techniques using microphone arrays for the same purpose, our algorithm is less complex and very accurate. It was shown that two concurrent sound sources could be perfectly localized at their intended 3D locations even in the anti-causal case where both sources share the same hemisphere around the humanoid's head. This is a remarkable improvement compared to the initially proposed adaptive MIMO approach.

The self-splitting competitive learning technique, mainly deployed in image processing, turned out to be very reliable for acoustical signal processing. It proved to be an intelligent tool to retrieve the exact cluster centers inside the feature space of the impinging sound signals, and consequently, to extract the 3D locations of the concurrent sound sources. After localization, the proposed sound source separation algorithm proved to be outperforming compared to other blind source separation methods solving the same determined problem under the same conditions.

For highly reverberant environments, a new algorithm using four microphones is presented. Bayesian information fusion is then used to increase the localization resolution in a three-dimensional reverberant environment. The algorithm was tested in simulations as well as in a household environment. Compared to existing techniques, the method is able to localize sound sources in three dimensions, under high reverberation conditions, with fewer sensors and higher accuracy.

Based on the simplicity of the presented new approach for sound source localization, the integration of audio with other modalities like haptic and vision becomes promising. This kind of integration will allow the multi-sensory telepresence system to improve the perceived degree of immersion for the human operator. Inspired by the human binaural hearing, this development is adding to the solution and attractiveness of the humanoid hearing technology, as humanoids share many characteristics with the human being, which is, after all, the most interesting object of scientific research.

# Appendix A

# Inner-outer Factorization Theorem Proof

**Theorem**: Let $W$ be a unitary matrix, and $Y$ be a uniformly bounded matrix which satisfies the following equality,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} Y & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} Y & B_0 \\ 0 & D_0 \end{bmatrix} W \tag{A.1}$$

such that $Y$ has a maximal dimension and $\ker(Y.) = 0$. Let

$$W = \begin{bmatrix} A_v & B_v \\ C_v & D_v \end{bmatrix}. \tag{A.2}$$

Then $\{A_v, B_v, C_v, D_v\}$ is an isometric realization for the sought inner factor $V(z)$, and $\{A, B, C_0, D_0\}$ is a realization for the outer factor $H_o(z)$. The proof of the above theorem is detailed in [33].

**Proof**: Since the $Y$ sought is such that $\text{kern}(Y.) = 0$, using $RQ$-factorization or SVD, we can always express it as

$$Y = V \begin{bmatrix} \sigma \\ 0 \end{bmatrix}, \tag{A.3}$$

in which $\sigma$ is square non-singular and $V$ is unitary.

We now let $\Delta$ be a block Schur eingenspace decomposition for the term $A - BD^{-1}C$,

$$\Delta = A - BD^{-1}C = V^{\star} \begin{bmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{bmatrix} V, \tag{A.4}$$

where $\delta_{11}$ collects the eigenvalues of $\Delta$ which are strictly outside the unit circle, thus, causing instability.

Given the state-space realization $\{A, B, C, D\}$, and assuming that $D$ is square invertible, we can write down the following schur factorization

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Delta & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}, \tag{A.5}$$

where $\Delta = A - BD^{-1}C$ is the Schur complement of D. We can now write (A.1) as

$$\begin{bmatrix} \Delta & 0 \\ C & D \end{bmatrix} \begin{bmatrix} Y & 0 \\ 0 & I \end{bmatrix} W^{\star} = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} Y & B_0 \\ 0 & D_0 \end{bmatrix}. \tag{A.6}$$

Looking at the second block column of this equation, we find

$$\begin{bmatrix} \Delta & 0 \\ C & D \end{bmatrix} \begin{bmatrix} Y & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} W_{21}^{\star} \\ W_{22}^{\star} \end{bmatrix} = \begin{bmatrix} Y \\ 0 \end{bmatrix}, \tag{A.7}$$

where $W_{21}^{\star} = A_v^{\star}$ and $W_{22}^{\star} = B_v^{\star}$.

From the last equation we can write, $\Delta Y = Y W_{21}^{-\star}$, or $W_{21}^{-\star} = Y^{+} \Delta Y$, where

$$Y^{+} = \begin{bmatrix} \sigma^{-1} & 0 \end{bmatrix} V^{\star}, \tag{A.8}$$

thus

$$W_{21}^{-\star} = \begin{bmatrix} \sigma^{-1} & 0 \end{bmatrix} V^{\star} \Delta V \begin{bmatrix} \sigma \\ 0 \end{bmatrix} \tag{A.9}$$

$$= \sigma^{-1} \delta_{11} \sigma. \tag{A.10}$$

It is important, at this point, to observe that since $W_{21} = A_v = \sigma^{\star} \delta_{11}^{-\star} \sigma^{-\star}$, the matrix, $\delta_{11}^{-\star}$, must have its eigenvalues strictly inside the unit disc, achieving thus the sought system stability.

Multiplying (A.7) with $Y^+$ we get

$$\begin{bmatrix} \sigma^{-1}\delta_{11}\sigma & 0 \\ CYD & \end{bmatrix} \begin{bmatrix} W_{21}^\star \\ W_{22}^\star \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix}. \tag{A.11}$$

Introducing

$$\beta = D^{-1}CV \begin{bmatrix} \delta_{11}^{-1} \\ 0 \end{bmatrix},$$

we find

$$\begin{bmatrix} W_{21}^\star \\ W_{22}^\star \end{bmatrix} = \begin{bmatrix} \sigma^{-1}\delta_{11}^{-1}\sigma & 0 \\ -\beta\sigma & D^{-1} \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix}. \tag{A.12}$$

Finally, setting $M = \sigma^{-\star}\sigma^{-1}$, we obtain the Lyapunov-Stein equation

$$M = \beta^\star\beta + \delta_{11}^{-\star} M \delta_{11}^{-1}. \tag{A.13}$$

Knowing that $\delta_{11}^{-1}$ has its eigenvalues strictly inside the unit disc of the complex plane, (A.13) must have a unique solution, and both $W_{21} = A_v$ and $W_{22} = B_v$ have also unique solutions,

$$\begin{bmatrix} W_{21}^\star & W_{22}^\star \end{bmatrix} = \begin{bmatrix} \sigma^\star\delta_{11}^{-\star}\sigma^{-\star} & -\sigma^\star\beta^\star \end{bmatrix}. \tag{A.14}$$

Once we find both $A_v$ and $B_v$ realizations corresponding to the inner-factor $V(z)$, we can proceed to compute $Y$ in (A.3) and use it in (A.1) to solve a system of four equations for the remaining inner-factor realizations $C_v$ and $D_v$ as well as the outer-factor, $H_o(z)$, realizations $C_0$ and $D_0$. Finally, the sought stable inverse, $H^{-1}(z) = H_o(z)^{-1}V^\star(z)$, is calculated, in which $H_o(z)^{-1}$ is singular and causal, and $V^\star(z)$ is anticausal and stable.

It should be noted that the outer-inner factorization computed using the Lyapunov-Stein equation, ensures a linear solution adequate for a fast real-time implementation, compared with other methods [139] which solve the same outer-inner factorization problem quadratically using the celebrated Riccati equation, and needlessly condition a problem which is already well-conditioned.

# Appendix B

# The State-Space Loewner Matrix

We will recapitulate the major steps involved in computing the $L$ matrix as detailed in [8]. To begin with, the generalized controllability matrix $N$ is first partitioned as $N = \begin{bmatrix} N_1 & N_2 \end{bmatrix}$ where $N_1 = (t_1 I - A)^{-1} B$. Define the term

$$\bar{N} = N_2 - \begin{bmatrix} N_1 & N_1 \dots N_1 \end{bmatrix} = NJ, \tag{B.1}$$

where

$$J = \begin{bmatrix} -I & -I & \dots & -I \\ I & 0 & \dots & 0 \\ 0 & I & \dots & 0 \\ 0 & 0 & \dots & I \end{bmatrix} \tag{B.2}$$

Define next

$$\begin{aligned} \tilde{N} &= N_2 \, diag[t_2 I, t_3 I, \dots, t_\delta I] - t_1 \begin{bmatrix} N_1, N_1, \dots, N_1 \end{bmatrix} \\ &= NJ_t, \end{aligned} \tag{B.3}$$

where

$$J_t = \begin{bmatrix} -t_1 I & -t_1 I & \dots & -t_1 I \\ t_2 I & 0 & \dots & 0 \\ 0 & t_3 I & \dots & 0 \\ 0 & 0 & \dots & t_\delta I \end{bmatrix} \tag{B.4}$$

Now that $\bar{N}$ and $\tilde{N}$ are defined, the state-space realizations $\{A, B, C, D\}$ which ensures a minimum degree for the interpolation transfer-function matrix $Y(x)$ can be computed according to the following equations,

$$A = \tilde{N}\bar{N}'(\bar{N}\bar{N}')^{-1} \tag{B.5}$$

$$B = (t_1 I - A)N \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{B.6}$$

$$C = \begin{bmatrix} I & 0 \dots & 0 \end{bmatrix} M(r_1 I - A) \tag{B.7}$$

$$D = Y(r_1) - C(r_1 I - A)^{-1} B \tag{B.8}$$

Note that $\bar{N}$ should have a full row rank for the inverse in (B.5) to exist. This $\{A, B, C, D\}$ realization ensures that the transfer-function $Y(x)$ interpolates the data and has least degree among interpolating transfer-function matrices [8].

# Appendix C

# List of Frequently Used Acronyms

| Acronym | Description |
|---|---|
| ADPF | Active Direction-Pass Filter |
| APV | Asymptotic Property Vector |
| BSS | Blind Source Separationt |
| CIPIC | Center for Image Processing and Integrated Computing |
| CPT | Conditional Probability Table |
| CPV | Center Property Vector |
| DLOS | Direct Line Of Sight |
| DOA | Direction of Arrival |
| DFT | Discrete Fourier Transform |
| DPV | Distant Property Vector |
| FFT | Fast Fourier Transform |
| FIR | Finite Impulse Response |
| HRIR | Head-Related Impulse Response |
| HRTF | Head-Related Transfer Function |

| **Acronym** | **Description** |
|---|---|
| ICA | Independent Component Analysis |
| IID | Interaural Intensity Difference |
| ILD | Interaural Level Difference |
| IPD | Interaural Phase Difference |
| ITD | Interaural Time Difference |
| ICA | Independent Component Analysis |
| KEMAR | Knowles Electronics Mannequin for Acoustic Research |
| MAMA | Minimum Audible Movement Angle |
| MIMO | Multiple-Input Multiple-Output |
| MISO | Multiple-Input Multiple-Output |
| MIT | Massachusetts Institute of Technology |
| OPTMC | One-Prototype-Take-Multiclusters |
| OPTOC | One-Prototype-Take-One-Cluster |
| SCA | Source Cancelation Algorithm |
| PRTF | Pinna-Related Transfer Function |
| SIMO | Single-Input Multiple-Output |
| SDR | Signal-to-Distortion Ratio |
| SIR | Signal-to-Interference Ratio |
| SNR | Signal-to-Noise Ratio |
| STFT | Short-Time Fourier Transform |
| 3D | Three-Dimensional |

# Bibliography


[1] P. Aarabi. The fusion of distributed microphone arrays for sound localization. *J. Acoust. Soc. Am.*, (4):338–347, 2003.

[2] P. Aarabi and B. Mungamuru. Scene reconstruction using distributed microphone arrays. In *Proc. International Conference on Multimedia and Expo*, pages 53–56, 2003.

[3] R. Aichner, H. Buchner, F. Yan, and W. Kellermann. Real-time convolutive blind source separation based on a broadband approach. In *Proc. Int. Symp. ICA*, pages 833–840, Granada, Spain, 2004.

[4] V. Algazi, R. Duda, R. Duraiswami, N. Gumerov, and Z. Tang. Approximating the head-related transfer function using simple geometric models of head and torso. *J. Acoust. Soc. Am.*, 112(5):2053–2064, 2002.

[5] V. Algazi, R. Duda, D. Thompson, and C. Avendano. Structural composition and decomposition of hrtfs. In *Proc. IEEE Workshop on Appl. of Sig. Proc. to Aud. and Acc.*, pages 103–106, New Paltz, NY, 2001.

[6] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The cipic hrtf database. In *Proc. 2001 IEEE Work. on Appl. of Sig. Proc. to Audio and Electroacoustics*, pages 21–24, New Paltz, NY, 2001. HRTF data sets are available at http://interface.cipic.ucdavis.edu/.

[7] J.B. Allen and D.A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, 1979.

[8] B.D.O. Anderson and A.C. Antoulas. Rational interpolation and state-variable realizations. *Linear Algebra and its Applications*, 137:479 – 509, 1990.

[9] S. Andersson, A. Handzel, V. Shah, and P. Krishnaprasad. Robot phonotaxis with dynamic sound-source localization. In *Proc. IEEE International Conference on Robotics and Automation*, pages 4833–4838, 2004.

[10] A.C. Antoulas and B.D.Q. Anderson. On the scalar rational interpolation problem. *IMA J.Math. Control and Inorm.*, 3:61–81, 1986.

[11] F. Avanzini, D. Rocchesso, and S. Serafin. Friction sounds for sensory substitution. In *Proc. Int. Conf. on Auditory Display*, volume 4, pages 1–8, 2004.

[12] D. W. Bateau. Listening with the naked ear. *S. J. Freedman, editor, Neuropsychology of Spatially Oriented Behavior. Dorsey Press, Homewood, IL. USA*, 1968.

[13] D. Begault, M. Godfroy, J. Miller, A. Roginska, M. Anderson, and E. Wenzel. Design and verification of headzap, a semi-automated hrir measurement system. In *Proc. 120th Convention of the Audio Engineering Society*, Paris, France, 2006. Product information can be found at http://ausim3d.com/products.

[14] D.R. Begault. 3-d sound for virtual reality and multimedia. Academic Press, 1994.

[15] B. Beliczynski, I. Kale, and G.D. Cain. Low-order modeling of head-related transfer functions using balanced model truncation. *IEEE Trans. Signal Processing*, 40(3):532–542, 1997.

[16] E. Ben-Reuven and Y. Singer. Discriminative binaural sound localization. In *Proc. Neural Information Processing Systems (NIPS)*, pages 1229–1236, 2002.

[17] L. Bernstein, C. Trahiotis, M. Akeroyd, and K. Hartung. Sensitivity to brief changes of interaural time and interaural intensity. *J. Acoust. Soc. Am.*, 109:1604–1615, 2001.

[18] S. Blackman and R. Popoli. Design and analysis of modern tracking systems(Book). *Norwood, MA: Artech House*, 1999.

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

[20] A. Blin, S. Araki, and S. Makino. A sparseness mixing matrix estimation (smme) solving the underdetermined bss for convolutive mixtures. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 85–88, Genua, Italy, 2004.

[21] S. E. Boehnke, S. E. Hall, and T. Merquadt. Detection of static and dynamic changes in interaural correlation. *J. Acoust. Soc. Am.*, 112:1617–1626, 2002.

[22] P. Bofill and M. Zibulevsky. Blind separation of more sources than mixtures using sparsity of their short-time fourier transform. In *Proc. Int. Workshop Independent Component Anal. Blind Signal Separation*, page 8792, Helsinki, Finland, June 2000.

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

[24] H. Buchner, R. Aichner, and W. Kellermann. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Transactions on Speech and Audio Processing*, 13(1):120–134, 2005.

[25] H. Buchner, R. Aichner, and W. Kellermann. A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Signal Processing*, 86:1260–1277, 2006.

[26] J. Merimaa C. Faller. Source localization in complex listening situations: selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Amer.*, 116:3075–3089, 2004.

[27] S. Carlile and D. Pralong. The location-dependent nature of perceptually salient features of the human head-related transfer functions. *J. Acoust. Soc. Am.*, 95(6):3445–3459, 1994.

[28] C. Cheng and G. Wakefield. Spatial frequency response surfaces (sfrs's): An alternative visualization and interpolation technique for head-related transfer functions (hrtf's). In *Proc. of the 16th Audio Eng. Soc. (AES) Int. Conf. on Spatial Sound Reproduction*, pages 961–964, Rovaniemi, Finland, 1999.

[29] 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.*, pages 1147–1152, USA, 2003.

[30] Y. Cohen and E. Knudsen. Representation of binaural spatial cues in field l of the barn owl forebrain. *J. Neurophysiol*, 79:879–890, 1998.

[31] H. Colburn and J. Latimer. Theory of binaural interaction based on auditory-nerve data, iii. joint dependence on inter-aural time and amplitude differences in discrimination and detection. *J. Acoust. Soc. Am.*, 64:95–106, 1978.

[32] J. Culling and H. Colburn. Binaural sluggishness in the perception of tone sequences and speech in noise. *J. Acoust. Soc. Am.*, 79:1939–1949, 1986.

[33] P. Dewilde and A. J. van der Veen. Inner-outer factorization and the inversion of locally finite systems of equations. *Linear Algebra and Its Applications*, 313:53–100, 2000.

[34] R. Duda and W. Martens. Range depend-ence of the response of a spherical head model. *J. Acoust. Soc. Am.*, 104(5):3048–3058, 1998.

[35] R. Duraiswami, L. Zhiyun, D. Zotkin, E. Grassi, and N. Gumerov. Plane-wave decomposition analysis for spherical microphone arrays. In *Proc. IEEE Wrok. on Appl. of Sig. Proc. to Audio and Acoustics (WASPAA)*, pages 150–153, 2005.

[36] O. Faugeras, editor. *Three Dimensional Computer Vision: A Geometric Viewpoint*. MA: MIT Press, Cambridge, 1993.

[37] H. Fisher and S. Freedman. The role of the pinna in auditory localization. *J. Audit. Res.*, 8:15–26, 1968.

[38] K. Fukudome, Y. Tashiro, K. Takenouchi, T. Samejima, and N. Ono. Development of the fast measurement system for the listener's own head-related impulse responses. *J. Acoust. Soc. Am.*, 120(5):3094(A), 2006.

[39] W. G. Gardner and K. D. Marting. Hrtf measurements of a kemar. *J. Acoust. Soc. Amer.*, 97(6):3907–3908, 1995.

[40] D. Grantham. Detection and discrimination of simulated motion of auditory targets in the horizontal plane. *J. Acoust. Soc. Am.*, 107:517–527, 2000.

[41] E. Grassi and S. Shamma. A biologically inspired, learning, sound localization algorithm. In *Proc. Conference on Information Sciences and Systems*, pages 344–348, 2001.

[42] S. Handel. Space is to time as vision is to audition: seductive but misleading. *Journal of Experimental Psychology: Human Perception and Performance*, (14):315–317, 1988.

[43] A. Handzel. Planar spherical diffraction-arrays: Linear sound localization algorithms. In *Proc. Fourth IEEE Workshop on Sensor Array and Multichannel Processing*, pages 655–658, 2006.

[44] A. Handzel and P. Krishnaprasad. Biomimetic sound-source localization. *Information Fusion (Special Issue on Robust Speech Processing)*, 5(2):131–140, 2004.

[45] K. Hartung, J. Braasch, and S. Sterbing. Comparison of different methods for the interpolation of head-related transfer functions. In *proc. of 6th Aud. Eng. Soc. (AES) Int. Conf. on Spatial Sound Reprod.*, pages 19–28, Rovaniemi, Finland., 1999.

[46] T. Horiuchi. "seeing" in the dark: neuromorphic vlsi modeling of bat echolocation. *IEEE Signal Processing Magazine*, pages 134–139, 2005.

[47] S. Hosoe, T. Nishino, K. Itou, and K. Takeda. Measurement of head-related transfer functions in the proximal region. In *Proc. Forum Acusticum*, pages 2539–2542, Budapest, Hungary, 2005.

[48] J. Hrnstein, M. Lopes, J. Santos-Victor, and F. Lacerda. Sound localization for humanoid robots - building audio-motor maps based on the hrtf. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 1170–1176, Beijing, China, October 2006.

[49] J. Huang, K. Kume, A. Saji, M. Nishihashi, T. Watanabe, and W. Martens. Robotic spatial sound localization and its 3-d sound human interface. In *Proc. of the IEEE First Int. Symp. on Cyber Worlds (CW.02)*, pages 191–197, Tokyo, Japan, November 2002.

[50] J. Huang, N. Ohnishi, X. Guo, and N. Sugie. A computational model of the precedence effect based on echo-avoidance. In *Proc. ASA/ASJ Third Joint Meeting*, page 635640, Dec. 1996.

[51] J. Huang, N. Ohnishi, and N. Sugie. Sound localization in reverberant environment based on the model ofthe precedence effect. *IEEE Transactions on Instrumentation and Measurement*, 1997.

[52] J. Huang, N. Ohnishi, and N. Sugie. Spatial localization of sound sources: Azimuth and elevation estimation. In *Proc. IEEE Instrumentation and Measurement Conference*, pages 330–333, USA, May 1998.

[53] R. E. Irie. Robust sound localization: an application of an auditory perception system for a humanoid robot. Master's thesis, Massachussets Institute of Technology, June 1995.

[54] L. Jeffress. A place theory of sound localization. *J. Comp. Phsiol. Psychol.*, 41:35–39, 1948.

[55] F. V. Jensen. *An Introduction to Bayesian Networks.* Springer Verlag, New York, 1996.

[56] P. Julian, A. Andreou, L. Riddle, S. Shamma, D. Goldberg, and G. Cauwenberghs. A comparative study of sound localization algorithms for energy aware sensor network nodes. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 51(4):640 – 648, 2004.

[57] C.H. Keller and T.T. Takahashi. Localization and identification of concurrent sounds in the owl's auditory space map. *Journal of Neuroscience*, 25(45):10446–10461, 2005.

[58] F. Keyrouz and K. Diepold. Artificial neural network based sound localization. Technical Report LDV-TR-M4-02, TU München, Lehrstuhl für Datenverarbeitung (LDV), 2006.

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

[60] 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)*, pages 663–665, Vancouver, Canada, 2006.

[61] F. Keyrouz and K. Diepold. Monaurale lokalisierung. Technical Report LDV-TR-M4-07, TU München, Lehrstuhl für Datenverarbeitung (LDV), 2006.

[62] F. Keyrouz and K. Diepold. A rational hrtf interpolation approach for fast rendering of moving sound. In *Proc. IEEE Digital Signal Processing (DSP) Workshop*, pages 222–226, Yellowstone, Whyoming, USA, 2006.

[63] F. Keyrouz and K. Diepold. Robotic sound detection: A novel human-based approach. In *Proc. Of the 2nd Int. Workshop on Human-Centered Robotic Systems (HCRS'06)*, pages 25–31, Munich, Germany, 2006.

[64] F. Keyrouz and K. Diepold. Sound localization techniques: Implementation and comparison. Technical Report LDV-TR-M4-01, TU München, Lehrstuhl für Datenverarbeitung (LDV), 2006.

[65] F. Keyrouz and K. Diepold. Tracking of fast moving sources using kalman filters. Technical Report LDV-TR-M4-03, TU München, Lehrstuhl für Datenverarbeitung (LDV), 2006.

[66] 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, MIT Press*, 15(2):509–522, 2007.

[67] F. Keyrouz and K. Diepold. Implementation of a real-time 3d sound rendering system. Technical Report LDV-TR-M4-06, TU München, Lehrstuhl für Datenverarbeitung (LDV), 2007.

[68] F. Keyrouz and K. Diepold. Robotic binaural and monaural information fusion using bayesian networks type of communication. In *Proc. 5th IEEE Int. Symp. on Intelligent Signal Processing*, pages 79–83, Madrid, Spain, October 2007. (Excellent top 10 paper).

[69] F. Keyrouz and K. Diepold. A new hrtf interpolation approach for fast synthesis of dynamic environmental interaction. *Audio Engineering Society (AES) Journal*, January 2008.

[70] F. Keyrouz and K. Diepold. Self-splitting competitive learning for binaural sound localization and separation in telerobotics. *International Journal of Information Technology and Intelligent Computing (ITIC), IEEE Computational Intelligence Society*, 1, 2008. (Accepted for publication).

[71] F. Keyrouz and K. Diepold. A novel biologically-inspired neural network solution for robotic 3d sound source sensing. *Soft Computing Journal, Springer*, February 2008. (Accepted for publication).

[72] F. Keyrouz, K. Diepold, and P. Dewilde. Robust 3d robotic sound localization using state-space hrtf inversion. In *proceedings of IEEE Int. Conf. on Robotics and Biomimetics (ROBIO)*, pages 245–250, Kunming, China, 2006. (Best Paper Award Nomination).

[73] F. Keyrouz, K. Diepold, and S. Keyrouz. High performance 3d sound localization for surveillance applications. In *Proc. of IEEE Int. Conference on Advanced Video and Signal Based Surveillance (AVSS)*, London, UK, 2007.

[74] F. Keyrouz, K. Diepold, and S. Keyrouz. Humanoid binaural sound tracking using kalman filtering and hrtfs. *Book of Robot Motion and Control: Lecture Notes in Control and Information Sciences (LNCIS), Springer Verlag, London*, pages 329–340, 2007.

[75] F. Keyrouz, K. Diepold, and S. Keyrouz. Kalman filtering for three dimensional sound tracking. In *Proc. 6th IEEE Int. Workshop on Robot Motion and Control*, pages 114–118, Poland, June 2007.

[76] F. Keyrouz and S. Keyrouz. Humanoid monaural sound localization using usupervised clustering. In *Proc. IEEE Int. Conference on Signal Processing and Communication (ICSPC07)*, Dubai, November 2007. (Accepted for publication).

[77] F. Keyrouz, F. Lazarno-blasco, and K. Diepold. Hierarchical fuzzy neural networks for robotic 3d sound source sensing. *Lecture Notes in Computer Science (LNCS), Springer Verlag*, 2007. (Accepted for publication).

[78] F. Keyrouz, F. Lazaro-blasco, and K. Diepold. Robotic sound detection based on hierarchical neural networks. In *Proc. IEEE intl. Symp. on Neural Networks (ISNN)*, China, 2007.

[79] F. Keyrouz, W. Maier, and K. Diepold. A novel humanoid binaural 3d sound localization and separation algorithm. In *Proc. IEEE-RAS International Conference on Humanoid Robots (Humanoids06)*, page 296301, Genua, Italy, December 2006.

[80] F. Keyrouz, W. Maier, and K. Diepold. Robotic binaural localization and separation of more than two concurrent sound sources. In *Proc. IEEE Int. Symposium on Signal Processing and its Applications (ISSPA)*, United Arab Emirates, 2007. 225-230.

[81] 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 Symposium on Computational Intelligence in Image and Signal Processing (CIISP)*, Hawaii, 2007. 340-345.

[82] 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)*, pages 341–344, Toulouse, France, May 2006.

[83] F. Keyrouz, A. Abou Saleh, and K. Diepold. Intelligent sound localization based on head-related transfer functions extraction. In *Proc. IEEE 3rd Int. Conf. on Intelligent Computer Comm. and Proc.(ICCP)*, pages 42–46, Romania, 2007.

[84] F. Keyrouz, A. Bou Saleh, and K. Diepold. A novel approach to monaural sound localization. In *Proc. 122nd Audio Engineering Society (AES) Convention*, Vienna, Austria, May 2007.

[85] F. Keyrouz, M. Usman, and K. Diepold. Real time 3d humanoid sound source localization in actual environments. In *IEEE 21st Canadian Conference on Electrical and Computer Engineering (CCECE08)*, Ontario, Canada, May 2008. (Accepted for publictation).

[86] F. Keyrouz, M. Usman, and K. Diepold. Real time humanoid sound source localization and tracking in a highly reverberant environment. In *Proc. 4th IEEE Int. Coll. on Signal Processing and its Applications (CSPA)*, Kuala Lumpur, Malaysia, March 2008. (Accepted for publictation).

[87] 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.*, 91(3):1637–1647, 1992.

[88] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[89] G. Kuhn. Model for the inter-aural time differences in the azimuthal plane. *J. Acoust. Soc. Am.*, 62:157–167, 1977.

[90] G. Metta L. Natale and G. Sandini. Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head. *Robotics and Autonomous Systems*, 39(2):87–106, 2002.

[91] Itakura Laboratory. Database of head related transfer functions, 1999. http://www.itakura.nuee.nagoya-u.ac.jp/.

[92] P. Lax and R. Phillips, editors. *Scarrering Theory.* Academic Press, NY, 1989.

[93] Paris France 2002 Listen HRTF Database, IRCAM. HRTF data sets are available at http://recherche.ircam.fr/equipes/salles/listen.

[94] C. Liu and B. Wheeler et al. Localization of multiple sound sources with two microphones. *Journal of the Acoustical Society of America*, 108(4):1888–1905, 2000.

[95] J. Mackenzie, J. Huopaniemi, V. Vlimki, and I. Kale. Low-order modeling of head-related transfer functions using balanced model truncation. *IEEE Signal Processing Letters*, 4(2):39–41, 1997.

[96] W. Martens. Principal components analysis and resynthesis of spectral cues to perceived direction. In *Proc. of the Int. Computer Music Conference*, pages 274–281, 1987.

[97] M. Matsumoto, S. Yamanaka, and M. Tohyama. Effect of arrival time correction on the accuracy of binaural impulse response interpolation. *Journal of the Audio Engineering Society*, 52:56–61, 2004.

[98] J. Middlebrooks. Narrow-band sound localization related to external ear acoustics. *J. Acoust. Soc. Am.*, 92(5):2607–2624, 1992.

[99] J. Middlebrooks and D. Green. Observations on a principal components analysis of head-related transfer functions. *J. Acoust. Soc. Amer.*, 92(1):597–599, 1992.

[100] W. Mills. Auditory localization. *Foundations of modern auditory theory*, II:303–348 (New York, NY, Academic Press), 1972.

[101] H. Moeller. Fundamentals of binaural technology. *Appl. Acoust.*, 36(3-4):171–218, 1992.

[102] M. Murase, S. Yamamoto, and J. Valin et al. Multiple moving speaker tracking by microphone array on mobile robot. In *Proc. European Conf. on Speech Communication and Technology (Interspeech)*, pages 249–252, September 2005.

[103] M. Murase, S. Yamamoto, J. Valin, and et al. Multiple moving speaker tracking by microphone array on mobile robot. In *Proc. European Conference on Speech Communication and Technology (Interspeech)*, volume 1, pages 143–145, 2005.

[104] J. Murray, H. Erwin, and S. Wermter. Recurrent neural network for sound-source motion tracking and prediction. In *Proc. IEEE Int. Joint Conf. on Neural Network (IJCNN)*, volume 4, pages 2232–2236, 2005.

[105] K. Nakadai, D. Matsuura, H. Okuno, and H. Kitano. Applying scattering theory to robot audition system. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1147–1152, Japan, September 2004.

[106] K. Nakadai, H. Okuno, and H. Kitano. Auditory fovea based speech separation and its application to dialog system. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1320–1325, Geneva, October 2002.

[107] K. Nakadai, H. G. Okuno, and H. Kitano. Realtime sound source localization and separation for robot audition. In *Proc. IEEE Int. Conf. on Spoken Language Processing*, pages 193–196, 2002.

[108] H. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano. Human-robot interaction through real-time auditory and visual multiple-talker tracking. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1402–1409, Hawaii, October 2001.

[109] J. Plogsties P. Minaar and F. Christensen. Directional resolution of head-related transfer functions required in binaural synthesis. *Journal of the Audio Engineering Society*, 53(10):919–929, 2005.

[110] S. Perrett and W. Noble. The effect of head rotations on vertical plane sound localization. *J. Audit. Res.*, 102(4):2325–2332, 1997.

[111] B. Radlovic, R. Williamson, and R. Kennedy. Equalization in an acoustic reverberant environment: Robustness result. In *Proc. IEEE Transactions on Speech and Audio Processing*, volume 8, page 311319, 2000.

[112] V. C. Raykara and R. Duraiswami. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *J. Acoust. Soc. Am.*, 118(1):364–374, 2005.

[113] J. Strutt (Lord Rayleigh). On our perception of sound direction. *Philosophical Magazine*, 13:214–232, 1907.

[114] K. Riederer. *HRTF analysis: Objective and subjective evaluation of measured head-related transfer functions*. Phd dissertation, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Espoo, Finland, 2005.

[115] S.M. Robeson. Spherical methods for spatial interpolation: review and evaluation. *Cartography and Geographic Information Systems*, 24(1):3–20, 1997.

[116] Y. Rui and D. Florencio. New direct approaches to robust sound source localization. In *Proc. of IEEE International Conf. on Multimedia Expo (ICME)*, pages 6–9, 2003.

[117] T. Shimoda, T. Nakashima, M. Kumon, R. Kohzawa, I. Mizumoto, and Z. Iwai. Spectral cues for robust sound localization with pinnae. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 386–391, China, October 2006.

[118] B. Shinn-Cunningham and A. Kulkarni. Recent developments in virtual auditory space. *Simon Carlile, Ed., Virtual Auditory space: Generation and Applications*, pages 185–243, (Austin, TX: R. G. Landes Company), 1996.

[119] B. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: Binaural cues for sources within reach of a listener. *J. Acoust. Soc. Am.*, 107(3):1627–1636, 1999.

[120] A. Steinhauser. The theory of binaural audition. *Phil. Mag.*, 7:261–274, 1877.

[121] L. Sbaiz T. Ajdler and M. Vetterli. The plenacoustic function on the circle with application to hrtf interpolation. In *proceedings of IEEE ICASSP*, 2005.

[122] S. Takane, D. Arai, T. Miyajima, K. Watanabe, Y. Suzuki, and T. Sone. A database of head-related transfer functions in whole directions on upper hemisphere. *Acoustical Science and Technology*, 23(2):160–162, 2002. HRTF data sets are available at the Suzuki Laboratory, Tohoku University, Japan, http://www.ais.riec.tohoku.ac.jp.

[123] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi. Three ring microphone array for 3d sound localization and separation for mobile robot audition. In *Proc. IEEE International Conference on Intelligent Robots and Systems*, pages 4172–4177, USA, 2005.

[124] I. Toshima and S. Aoki. The effect of head movement on sound localization in an acoustical telepresence robot: TeleHead. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 872–877, China, October 2006.

[125] I. Toshima, S. Aoki, and T. Hirahara. An acoustical tele-presence robot: TeleHead II. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2105–2110, Japan, September 2004.

[126] J. Valin, F. Michaud, J. Rouat, and D. Ltourneau. Robust sound source localization using a microphone array on a mobile robot. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 1228– 1233, Oct. 2003.

[127] J. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS))*, pages 1033–1038, 2004.

[128] 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*, pages 1228–1233, Saitama, Japan, 2003.

[129] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[130] A. Vljame, P. Larsson, D. Vstfjll, and M. Kleiner. Auditory presence, individualized head-related transfer functions, and illusory ego-motion in virtual environments. In *Proc. Presence*, pages 141–147, Valencia, Spain, Oct. 2004.

[131] H. Wallach. On sound localization. *J. Acoust. Soc. Am.*, 10(4):270–274, 1939.

[132] G. Welch and G. Bishop. An intoduction to the kalman filter. Tr95-041 tech. rep., Univ. of North Carolina, Department of Computer Science, 1995.

[133] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, 94:111–123, 1993.

[134] F. Wightman and D. Kistler. Headphone simulation of free-field listening: Stimulus synthesis. *J. Acoust. Soc. Am.*, 1989.

[135] F. Wightman and D. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, 91(3):1648–1661, 1992.

[136] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Krner. A probabilistic model for binaural sound localization. *IEEE Transactions on Systems, Man, and Cybernetics*, 36(65):982–994, 2006.

[137] K. Wilson and T. Darrel. Improving audio source localization by learning the precedence effect. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 1125–1128, 2005.

[138] S. Winter, W. Kellermann, H. Sawada, and S. Makino. Map based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization. *EURASIP Journal on Advances in Signal Processing*, page Article ID 24717, 2007.

[139] K. Yamada and K. Watanabe. Inner-outer factorization for the discrete-time strictly proper systems. In *proceedings of IEEE 35th Conf. on Decison and Control*, pages 1491–1492, 1996.

[140] Y.-J. Zhang and Z.-Q. Liu. Self-splitting competitive learning: A new on-line clustering paradigm. *IEEE Transactions on Neural Networks*, 13(2):369–380, 2002.

[141] D. Zotkin, R. Duraiswami, E. Grassi, and N. Gumerov. Fast head related transfer function measurement via reciprocity. *J. Acoust. Soc. Am.*, 120(4):2202–2215, 2006.