Fakheredine Keyrouz*
Klaus Diepold
Data Processing Institute
Technische Universität München
80290 Munich, Germany

Binaural Source Localization and Spatial Audio Reproduction for Telepresence Applications

Abstract

Telepresence is generally described as the feeling of being immersed in a remote environment, be it virtual or real. A multimodal telepresence environment, equipped with modalities such as vision, audition, and haptic, improves immersion and augments the overall perceptual presence. The present work focuses on acoustic telepresence at both the teleoperator and operator sites. On the teleoperator side, we build a novel binaural sound source localizer using generic Head Related Transfer Functions (HRTFs). This new localizer provides estimates for the direction of a single sound source given in terms of azimuth and elevation angles in free space by using only two microphones. It also uses an algorithm that is efficient compared to the currently known algorithms used in similar localization processes. On the operator side, the paper addresses the problem of spatially interpolating HRTFs for densely sampled high-fidelity 3D sound synthesis. In our telepresence application scenario the synthesized 3D sound is presented to the operator over headphones and shall achieve a high-fidelity acoustic immersion. Using measured HRTF data, we create interpolated HRTFs between the existing functions using a matrix-valued interpolation function. The comparison with existing interpolation methods reveals that our new method offers superior performance and is capable of achieving high-fidelity reconstructions of HRTFs.

1 Introduction

Telepresence systems aim at supplying the senses of a human operator with stimuli that 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. The perceptual world that the operator is experiencing is built up of sensory data that is provided by a teleoperator, that is, a robot, located at a remote site. At the local operator site, a human operator is interacting with a multimodal 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

Presence, *Vol. 16, No. 5, October 2007,* 509–522

*Correspondence to keyrouz@tum.de.

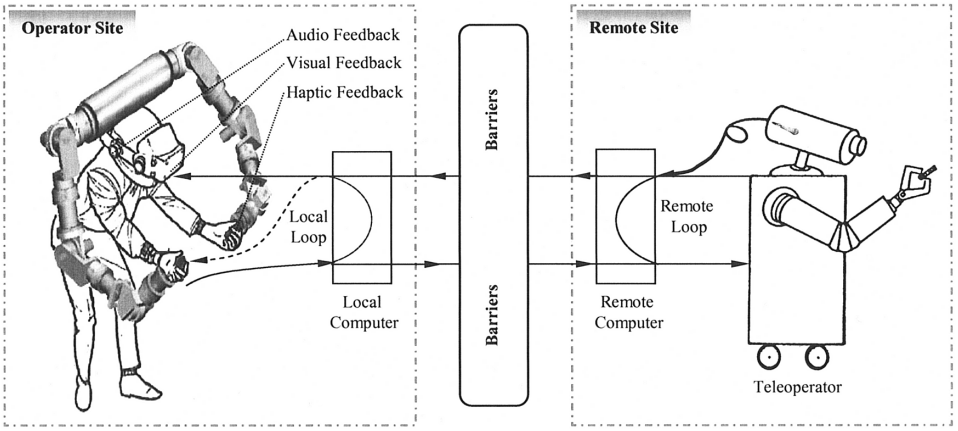

Figure 1. *A multimodal telepresence system.*

audio with other modalities, such as 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 (1988) 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 (Avanzini, Rocchesso, & Serafin, 2004; Altinsoy, 2003). Furthermore, while vision is a sense that is directed, audio is an undirected sense that helps the human perceive and locate audible events that happen outside his field of vision. This feature is important also in the case of multiple teleoperators acting at the remote site, since it enables the operator to receive warnings and cues of activities not visible to him (see Figure 1).

1.1 Acoustic Telepresence

The present work focuses on the auditory modality. In a general acoustic teleoperation we consider both the teleoperator site (binaural source localization), as well as the operator site (spatial sound synthesis). In the context of acoustic telepresence, as shown in Figure 2, 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

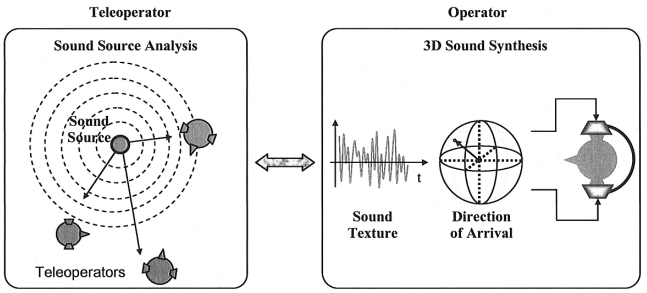

Figure 2. *Acoustic telepresence scenario.*

identify the sound source location in the surrounding environment. 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 direction of the sound source as located at the teleoperator site. Note 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.

1.2 Binaural Hearing and Head Related Transfer Functions

When placed in a free sound field, a listener will obstruct an incoming sound wave. The listener's ears, head, and body will produce reflections and hence influence the sound wave propagation. This influence can be represented as a linear filtering operation of the original sound signal by means of the so-called Head Related Transfer Functions (HRTFs). In the general definition of the HRTF, all linear properties of the sound transmission are included. In order to synthesize a spatialized sound impression, a sound texture is filtered with a HRTF, which describes the transfer of a sound arriving from a given 3D direction of a source to the ear canals of a listener. To each 3D direction a corresponding HRTF is associated, which is given as a one-input two-output linear transfer function. All proposed descriptors of localization cues, such as inter-aural time difference in arrival-time (ITD), the inter-aural phase difference (IPD), the inter-aural level difference or intensity difference (ILD/IID) as well as monaural cues, are contained in the HRTFs. Therefore, those cues can be derived from the HRTFs, whereas the opposite is not generally the case (Blauert, 1997).

1.3 Binaural Spatial Sound Synthesis

The use of HRTFs for binaural synthesis of spatial audio is well known. On the human operator site, the success of binaural synthesis and the quality of the generated spatial sound strongly depend 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 sparse spatial grid of HRTFs is recorded. In order to represent all possible directions of the sound source with respect to the listener, a discrete number of HRTFs is measured and stored. To compute additional functions at intermediate grid points an interpolation technique is used. We present a new interpolation method to construct a rational interpolation function which, given two spatially neighboring measured HRTFs, can accurately interpolate a number of HRTFs in between. This way, we increase the fidelity of the 3D sound synthesis while avoiding noticeable artifacts due to switching between different HRTFs. We recapitulate first the work done at the teleoperator site before turning our attention toward the human operator part of the telepresence operation.

2 Binaural Sound Source Localization

In this section, we propose a binaural sound source localization technique, which is based on using only two small microphones placed inside the ear canal of a robot dummy head. The head is equipped with artificial ears and is mounted on a torso. In contrast to existing sound source localization methods, we employ a matched filtering approach using the HRTFs applied to the signals collected by the two microphones. This setup proves to be very noise-tolerant and is able to localize sound sources in free space with high precision (Keyrouz, Naous, & Diepold, 2006). Note that HRTFs so far have been mainly used for synthesis of spatial sound, while we are also using HRTFs for sound source localization.

2.1 State of the Art

For the problem of localizing the spatial position of a sound source, a number of models have already been proposed (Handzel, 2005; Nakashima & Mukai, 2005; Rui & Florencio, 2003). 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 (Duraiswami, Zhiyun, Zotkin, Grassi, & Gumerov, 2005). Adaptive neural network structures have also been proposed to self-adjust a sound localization model to particular environments (Murray, Erwin, & Wermter, 2005). These structures disregard the head and pinnae, and create a sort of scanning or beamforming system that can focus on the main source and atten-

uate reflections and 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 (Grassi & Shamma, 2001). Obviously, the difficulty with this approach is that neurophysiologists do not completely understand how living organisms localize sounds. For instance, the question of what primitive mammals such as bats experience and how they process sound with only two ears and a pea-sized brain remains a major mystery (Horiuchi, 2005).

Recently, there has been progress in sound localization using microphone arrays (Jahromi & Aarabi, 2005; Wang, Ivanov, & Aarabi, 2004). In Handzel and Krishnaprasad (2002), an algorithm was proposed that determines the direction of arrival of sound by devising two curves, the acoustical phase difference and the intensity level difference between two microphones as functions of the measured frequency. These curves are then weighed 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 wastes time distinguishing between the two hemispheres. The algorithm is limited to the bandwidth of the source and its performance deteriorates in the presence of acoustic and electronic noise.

2.2 Dummy Head HRTFs

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

KEMAR (Knowles Electronics Manikin for Acoustic Research) is a standard manikin based on common human anthropomorphic data and designed for making HRTF measurements (Gardner & Martin, 1995). The research group at the MIT Media Lab has made extensive measurements using KEMAR. They provide a dataset 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. We investigate three techniques for reducing the length of the HRTF, two FIR and one IIR, which are applied to the KEMAR dataset, and which lead to a significant reduction in the size of the measured HRTF dataset. Using the reduced dataset, we present a novel approach to localize sound sources using only two microphones in a real environment.

2.3 Model Reduction Techniques

The KEMAR dataset contains the impulse responses of the actual measured HRTF filters. The 512 samples of each HRTF measurement can directly be considered to be the coefficients of an FIR representation of the filter. However, for real-time processing FIR filters of this order are computationally expensive. Moreover, the dataset is to be used to perform localization of sound sources and to account for head movements, which implies that the dataset has to be stored to allow for fast switching between HRTFs. Using the 512 samples slows down the localization process and does not offer memory savings. The original KEMAR HRTFs containing the 512 coefficients of the FIR filter will be denoted by H_{512}^{FIR}.

2.3.1 Diffuse-Field Equalization. Our goal is to shorten the length of the original filters H_{512}^{FIR} in order to reduce the computational burden for convolution, while preserving the main characteristics of the measured impulse responses. To this end, we adopt the algorithm proposed by Møller (1992) for a diffuse-field equalization (DFE). In DFE, a reference spectrum is derived by power-averaging all HRTFs from each ear and taking the square root of this average spectrum. Diffuse-field equalized HRTFs are obtained by deconvolving the original by the diffuse-field reference HRTF of that ear. This leads to the fact that the factors that are not incident-angle dependent, such as ear canal resonance, are removed. The DFE is achieved according to the following steps:

- Remove the initial time delay from the beginning of the measured impulse responses, which typically has a duration of about 10–15 samples.
- Remove features from modeling that are independent of the incident angle, such as ear canal resonance, and loudspeaker and microphone responses (Møller, 1992).
- Smooth the magnitude response using a critical-band auditory smoothing technique (Mackenzie, Huopaniemi, Välimäki, & Kale, 1997).
- Construct a minimum-phase filter, thus ensuring stability for the final filter and its inverse.

This way we shorten the length of the FIR representation of the original KEMAR HRTFs from 512 to 128 coefficients. Figure 3 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 H_{128}^{FIR} and H_{512}^{FIR} models. For the 710 impulse responses that we modeled, SERs were in the range of 20–37 dB with an average of 30 dB.

2.3.2 Balanced Model Truncation. In order to examine to which extent the HRTF can be reduced while still preserving the characteristic information that

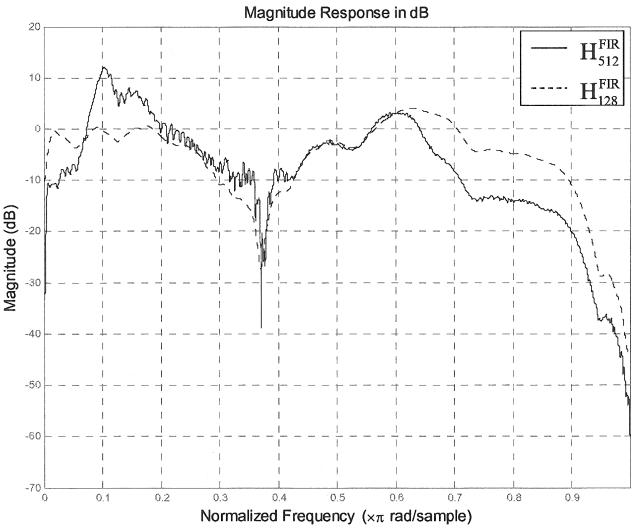

Figure 3. *Magnitude response of the original 512-FIR (solid) and the reduced 128-FIR (dashed) of an HRTF (left ear, 0° azimuth).*

makes it unique, we reduce the previously derived diffused-field HRTF dataset further by adopting the balanced model truncation (BMT) technique to design a low-order IIR filter model of the HRTF from a high-order FIR filter response H_{128}^{FIR}. A detailed description of the BMT technique is given in Beliczynski, Kale, and Cain (1992). However, a brief outline will be presented here.

For applying BMT we determine a linear time-invariant state-space system, which realizes the filter H_{128}^{FIR}. We start using the 128-coefficient FIR filter H_{128}^{FIR}. The transfer function of this filter can be written as: $F(z) = c_0 + c_1 \cdot z^{-1} + c_2 \cdot z^{-2} + \ldots + c_n \cdot z^{-n}$, where $n = 127$. This filter can be represented as state-space difference equations:

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

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

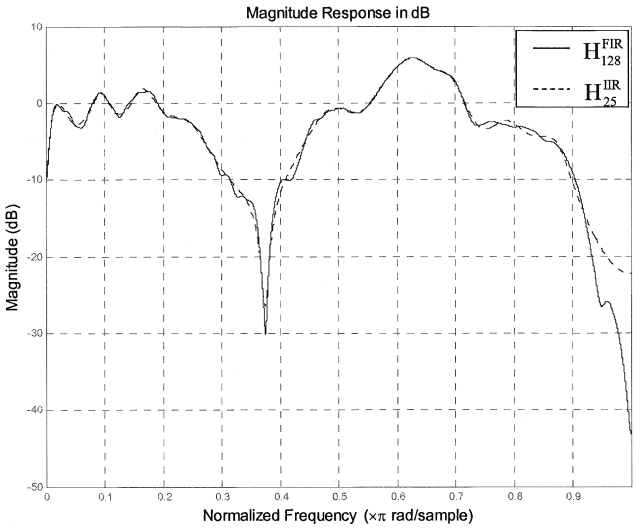

Figure 4. *Magnitude response of the 128-FIR (solid) and the reduced 25-IIR (dashed) of an HRTF (left ear, 0° azimuth).*

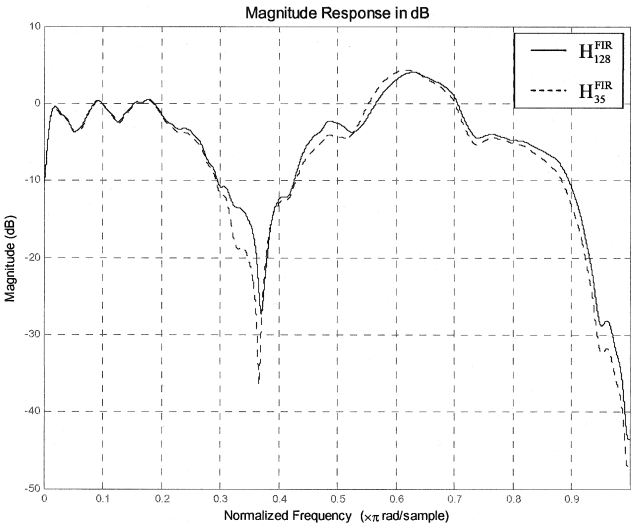

Figure 5. *Magnitude response of the 128-FIR (solid) and the reduced 35-FIR (dashed) of an HRTF (left ear, 5° azimuth).*

Singular Values (HSV) of the system. Thus, the balanced system can be divided into two subsystems: 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 4 shows the BMT-reduced IIR filter representation ($m = 25$) in comparison to the FIR ($n = 128$) for one example of an 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.

2.3.3 Principal Component Analysis. As an alternative to the previous BMT approach a Principal Component Analysis (PCA) is used to reduce the number of samples required to represent each 128-sample diffuse-field equalized HRTF. A thorough description of the PCA technique in modeling HRTFs is available in Kistler and Wightman (1992). The PCA aims at minimizing the amount of storage space needed for the HRTF dataset by selecting m representatives from the whole dataset. Figure 5 shows for a characteristic example that the PCA-reduced FIR ($m = 35$), represented as H_{128}^{FIR}, follows the general trend of the FIR ($n = 128$) filter.

For the difference between both FIR models, SERs were found to fall in the range of 23–37 dB with an average of 30 dB.

2.4 Proposed Sound Source Localization Technique

Our proposed approach for sound source localization starts by finding the inverse filters for the whole reduced HRTF dataset in an offline process, and then saving these inverses for later use. The inverse filter is directly available by simply exchanging the values of the numerator and the denominator, that is, for the FIR filter

$$F(z) = c_0 + c_1 \cdot z^{-1} + c_2 \cdot z^{-2} + \ldots + c_n \cdot z^{-n},$$

the inverse filter would be

$$G(z) = \frac{1}{F(z)}.$$

Similarly, the same concept applies for the IIR filter.

Our sound source localization scenario is as follows: multiple sound sources are generating signals; these signals are filtered with a corresponding HRTF, depending

on the location of the sound source. In order to ensure rapid localization of multiple sources, a small part of the filtered signal is considered (about 350 ms). This signal part is then convolved with the available reduced HRTF set. Since the original signals arrive in pairs, that is, one for the left ear and the other for the right ear, we calculate the correlation factor between the two resulting outputs. Basically, the correlation factor should yield a value that is close to one, since the two signals are supposed to be almost the same. Therefore, we base our localization on the obtained maximum for the correlation factor c. Moreover, to ensure 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 following is the flow of the algorithm:

$y_L(t)$ = received signal at left ear
$y_R(t)$ = received signal at right ear
n = number of HRTFs in the dataset
for $i = 1{:}n$
$x_L^{(i)}(t) = y_L(t) * H_L^{-1(i)}$;
$x_R^{(i)}(t) = y_R(t) * H_R^{-1(i)}$;
$c^{(i)} = corr(x_L^{(i)}(t), x_R^{(i)}(t))$
$d^{(i)} = \sum_m (x_L^{(i)}(t) - x_R^{(i)}(t))^2$
% m = degree of reduced HRTF
end

The above localization technique is further illustrated in Figure 6. Note that this sound source localization algorithm uses the previously described reduced HRTF models. The received signal S_R, inside the right ear, is transformed to the frequency domain, where it is divided by every HRTF inverse, denoted by $H_{R1} \cdots H_{R710}$. The same process is repeated for the left ear sound signal, S_L. The division output for every HRTF pair is cross correlated. The pair yielding maximum cross-correlation is selected to be the one corresponding to the target sound source location.

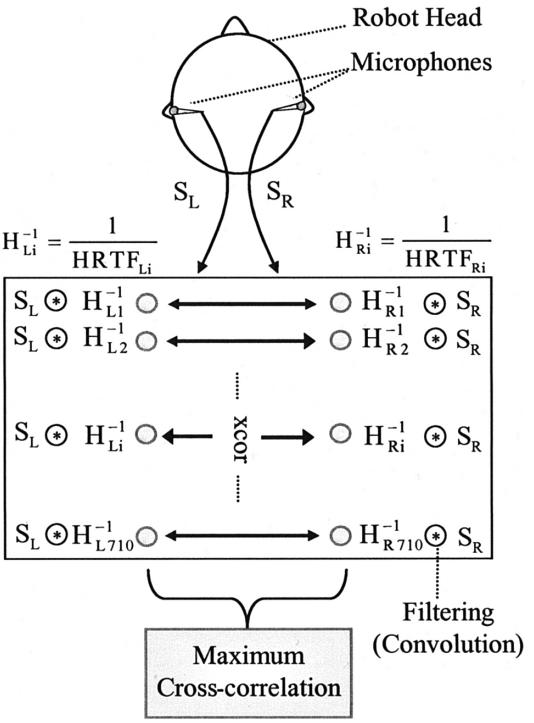

Figure 6. *Free space sound detection using only two microphones.*

2.5 Simulation and Experimental Results

The simulation test consisted of having a mono 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 KEMAR HRTFs were used corresponding to 100 different random source locations in the 3D space. The reduction techniques, namely, diffused-field equalization, balanced model truncation, and principle component analysis, were used to create three different reduced models of the original HRTFs. The performance of each of these models is illustrated in Figure 7. This figure shows the percentage of correct localization versus the length of the HRTF in samples.

Using the diffuse-field equalized 128 samples HRTF set, the simulated percentage of correct localization was around 96%. This means that out of 100 locations, 96 were detected by our algorithm, 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 data-

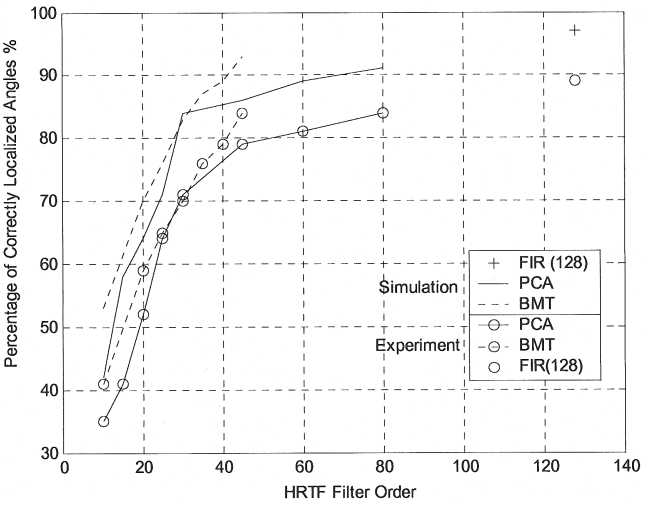

Figure 7. *Percentage of correct localization using reduced HRTFs.*

set being represented by 10 to 80 samples. Most significantly, it was observed that, starting with 30 samples PCA-reduced and 35 BMT-reduced HRTFs, all the falsely localized angles fall within 5° and 15°, with an average angular distance of 8°, to the target sound source locations.

On the other hand, 100 binaural recordings from different directions were obtained using a dummy head and torso with two artificial ears in a reverberant room. The microphones were placed near the eardrums at a distance of 26 mm away from the ear's opening. The recorded sound signals, also containing external and electronic noise, were used as inputs to our localization algorithm. The localization accuracy is depicted by the circled lines of Figure 7. The maximum accuracy observed is 89% and corresponds to a filter order of 128. The noticeable drop of accuracy is the result of the room reverberations, internal equipment noise, and is mainly caused by the differences between the dummy manikin model used in the experiment and the KEMAR model used to obtain the HRTF dataset. Using more than 30 samples PCA-reduced and 35 BMT-reduced HRTFs, the falsely localized angles fall within 5° and 35°, with an average distance of 10°, to the target sound source locations.

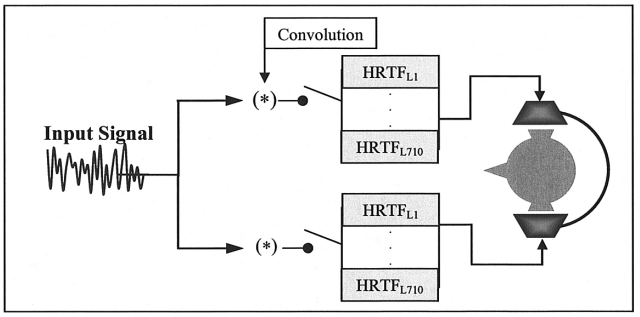

Figure 8. *Sound synthesis using HRTFs.*

3 3D Sound Synthesis

The direction of the 3D sound source at the teleoperator site, which has been determined according to our above-mentioned algorithm, is transmitted to the operator side. Optionally we can send along the original mono sound signal. Binaural signals are generated electronically by convolving an audio signal, either the original sound or some artificial sound texture with a measured set of HRTFs; see Figure 8. Since the process undertaken to measure the HRTF filters, which constitute the cornerstone of binaural synthesis, is complex and time-consuming, only a sparse spatial grid of HRTFs is recorded. Spatial interpolation techniques are used to increase the resolution of the grid by computing HRTFs corresponding to 3D positions that are located between the recorded functions. The need for interpolating HRTFs becomes more important when we address the problem of moving sound sources where we need to switch between different HRTFs or when the perceived direction of the sound changes due to the listener's head movements. It is well known that in real life head turns are often used unconsciously, for example, to resolve front/back confusion. The achievable immersion is dependent on the capability to accomplish this switching in a smooth way without creating audible artifacts.

The position and orientation of the listener's head can be tracked by so called head trackers, and the binaural synthesis can react to these changes by continuously updating the signal processing task, that is by fetching

the correct HRTF corresponding to these changes, or interpolating it in case it is not found. 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 (Minnaar, Plogsties, & Christensen, 2005).

Dynamic synthesis dictates the availability of HRTFs at any point on the 3D sphere surrounding the moving listener's head. Only then can localization errors such as front/back confusion be reduced, and cone-of-confusion errors can be well resolved. The term cone-of-confusion refers to a whole cone of points in three dimensional space that would produce the same ITDs and ILDs, causing, thus, ambiguities about the location they are generated from.

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

3.1 Previous Interpolation Methods

The bilinear method (Begault, 2004) 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 a weighted mean of the measured binaural responses and associated with the two adjacent points to the desired point. The discrete Fourier Transform (DFT) (Matsumoto, Yamanaka, & Tohyama, 2004) is used to interpolate binaural impulse responses in the time domain. The method inserts all HRTFs columnwise 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 (Ajdler & Vetterli, 2005) and the mathematical spline function (Robeson, 1997) have also been used to interpolate HRTFs.

In Matsumoto et al. (2004), the authors showed that with a moving sound source, the distance between the source and a listener varies as the source moves. They 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 (DFT), and third order spline function. The initial time-delay equalization was demonstrated to increase the accuracy of all proposed interpolation techniques.

In this section, 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 interpolations methods.

3.2 Rational Interpolation

The rational interpolation problem was first solved for the scalar transfer function case (Antoulas & Anderson, 1986). 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 Anderson and Antoulas (1990). We will recapitulate some of the key properties of the scalar rational interpolation problem, before tackling the state-space rational interpolation problem.

3.2.1 Scalar Rational Interpolation. Consider the array of points $P := \{(x_i, y_i),\ i = 1, \ldots, n\}$, with $x_i \neq x_j$ for $i \neq j$, and $x_i, y_j \in \mathbb{C}$. The fundamental rational interpolation problem is to parameterize all rational functions

$$y(x) = \frac{b(x)}{a(x)} \tag{2}$$

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 $y(x)$ is determined by

$$\sum_{i=1}^{n} c_i \frac{y(x) - y_i}{x - x_i} = 0, \quad c_i \neq 0. \tag{3}$$

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 rational degree of the interpolation function y. One way to do this is to consider a summation as in Equation 3 containing only $q < n$ summands, for any set of nonzero 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 $\mathbf{c} := [c_1, \ldots, c_q]^T$; in order for the remaining $n - q$ points to be interpolated, $\mathbf{c}$ must satisfy

$$\sum_{i=1}^{q} c_i \frac{y_{q+j} - y_i}{x_{q+j} - x_i} = 0, \quad i = 1, 2, \cdots, n - q, \tag{4}$$

or in matrix form

$$L \cdot \mathbf{c} = 0, \tag{5}$$

where L is a Loewner or *divided differences* matrix of dimension $(n - q) \times q$, 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's characteristics is found in Anderson and Antoulas (1990).

The interpolation problem now boils down to determining the $\mathbf{c}$ vector such that Equation 5 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

$$b(x, \mathbf{c}) = \sum_{i}^{n} c_i y_i \prod_{j \neq i} (x - x_j),$$
$$a(x, \mathbf{c}) = \sum_{i}^{n} c_i \prod_{j \neq i} (x - x_j). \tag{6}$$

For the proof of Equation 6 refer to Antoulas and Anderson (1986).

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.

3.2.2 State-Space Rational Interpolation. Let $\Upsilon(x)$ be a matrix-valued transfer-function with a minimal state-space realization $\{A, B, C, D\}$ of the form

$$\Upsilon(x) = C^*(xI - A)^{-1}B + D. \tag{7}$$

Consider the array of points $P := \{(x_i, y_i)/y_i = \Upsilon(x_i)$ for $i = 1, \ldots, n\}$, with $x_i \neq x_j$, and $x_i, y_i \in \mathbb{C}$. Suppose we now 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 $R = \{r_1, r_2, \ldots, r_\mathrm{j}\}$, with $r_i \neq r_j$ for $i \neq j$, be the row set and, 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 $\Upsilon(x)$ 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 $\Upsilon(x)$ is factored into a product of two matrices M and N with column and row rank q respectively,

$$-L = \begin{bmatrix} C^*(r_1 I - A)^{-1} \\ C^*(r_2 I - A)^{-1} \\ \vdots \\ C^*(r_\gamma I - A)^{-1} \end{bmatrix} \times [(t_1 I - A)^{-1}B(t_2 I - A)^{-1}B \cdots (t_\delta I - A)^{-1}B] = MN. \tag{8}$$

This is a state-space representation of the Loewner matrix. Similar to the Hankel matrix, the Loewner matrix is factored into a product of generalized controllability and observability matrices. The main strategy now is to find the Loewner matrix L, that is, to compute a realization $\{A, B, C, D\}$ such that M and N are the generalized observability and controllability matrices. We will recapitulate the major steps involved in computing the L matrix as detailed in Anderson and Antoulas (1990).

To begin with, the generalized controllability matrix N is first partitioned as $N = [N_1\ N_2]$ where $N_1 = (t_1 I - A)^{-1}B$. Define the term

$$\bar{N} = N_2 - [N_1 \quad N_1 \quad \cdots \quad N_1] = NJ \tag{9}$$

where

$$J = \begin{bmatrix} -I & -I & \cdots & -I \\ I & 0 & \cdots & 0 \\ 0 & I & \cdots & 0 \\ 0 & 0 & \cdots & I \end{bmatrix} \tag{10}$$

Define next

$$\begin{aligned} \tilde{N} &= N_2\, diag[t_2 I \quad t_3 I \quad \cdots \quad t_\delta I] \\ &\quad - [N_1 \quad N_1 \quad \cdots \quad N_1] \\ &= NJ_t \end{aligned} \tag{11}$$

where

$$J_t = \begin{bmatrix} -t_1 I & -t_1 I & \cdots & -t_1 I \\ t_2 I & 0 & \cdots & 0 \\ 0 & t_3 I & \cdots & 0 \\ 0 & 0 & \cdots & t_\delta I \end{bmatrix} \tag{12}$$

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

$$\begin{aligned} A &= \tilde{N}\bar{N}'(\bar{N}\bar{N}')^{-1} \\ B &= (t_1 \mathrm{I} - A)N \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ C &= [I \quad 0 \quad \cdots \quad 0]M(r_1 I - A) \\ D &= \Upsilon(r_1) - C(r_1 I - A)^{-1}B \end{aligned} \tag{13}$$

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

We can now formulate the rational interpolation problem to fit our aim of interpolating the HRTFs.

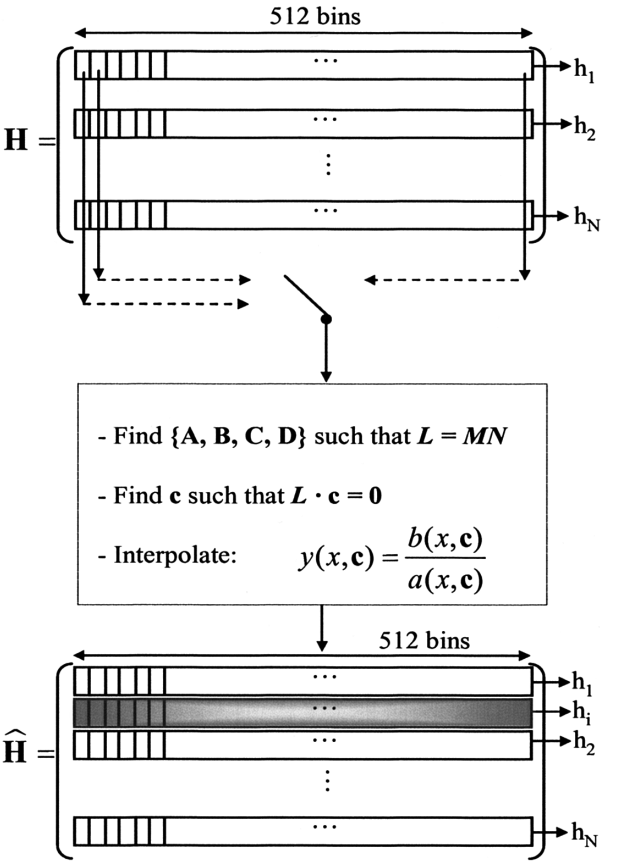

Figure 9. *The interpolation process.*

3.3 HRTF Interpolation Framework

In our investigation we use KEMAR HRTFs of length 512 samples measured at 44.1 kHz every 5° in the horizontal plane, a total of 72 locations. Figure 9 illustrates the interpolation procedure. After stacking all binaural responses (time-domain) as the rows of a common matrix **H**, the interpolation algorithm reads the matrix columnwise one column at a time.

Every column in the **H** matrix contains a number of 72 samples taken from every binaural impulse response. Every column can be written as $\Upsilon(z) = c_0 + c_1 \cdot z^{-1} + c_2 \cdot z^{-2} + \ldots + c_n \cdot z^{-n}$ where $n = 71$. This filter possesses a minimal state-space realization $\{A, B, C, D\}$ having the form of Equation 7.

We follow the state-space rational interpolation described in 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 *L* according to Equation 8. We then compute the vector **c** of interpolating coefficients following the Loewner matrix equation, $L \cdot \mathbf{c} = 0$. Now that the vector of interpolating coefficients is computed, we can directly obtain the desired interpolation function $y(x, \mathbf{c}) = \frac{b(x, \mathbf{c})}{a(x, \mathbf{c})}$, where $a(x, \mathbf{c})$ and $b(x, \mathbf{c})$ are defined in Equation 4. Using the $y(x, \mathbf{c})$ function for noninteger values of the variable x we are able to interpolate values for the impulse responses at new angles between 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 72 rows, the resulting matrix **H** is composed of rows that include the interpolated binaural responses.

Like the DFT method, the rational interpolation method uses the responses at all azimuths for the determination of the response to be interpolated. In the bilinear method, however, the binaural response to be interpolated is determined based only on two adjacent responses.

3.4 Discussion of Results

To verify the performance of the rational interpolation method, we use 72 measurements for the left ear. We then remove one of the HRTFs. The now-missing HRTF is reinserted using the interpolation technique described in Section 3.3. The interpolation result is compared with the corresponding available measurement and the Signal-to-Distortion Ratio (SDR) is computed according to

$$\mathrm{SDR} = 10\log \frac{\sum_{n=0}^{N_h-1} h_L^2(n)}{\sum_{n=0}^{N_h-1} [h_L(n) - \hat{h}_L(n)]^2} \qquad (14)$$

where $h_L(n)$ denotes a measured HRTF at a certain azimuth, for the left ear. The symbol $\hat{h}_L(n)$ denotes an interpolated HRTF at the same azimuth, and N_h is the HRTF length, for example, $N_h = 512$.

To evaluate our Rational State-Space (RSS) method, and find out improvements over previous work done on the same topic, we compare our interpolation results with three well-known HRTF interpolation techniques, namely the bilinear, DFT, and spline using time correction. The rational interpolation in Figure 10 shows comparable performance from 0° to 45° azimuths and from 345° to 360°. For the remaining azimuth range, the SDR value for the rational method is higher than that for the other three methods. The SDR value at 110° is 32 dB compared with an average of −4 dB for the other methods, that is, achieving a total of 32 dB gain. The SDR value reaches 72 dB at 270°, that is, a total of 42 dB gain. Furthermore, it should be noted that even more gain can be achieved if we use the time correction method along with the presented rational interpolation method.

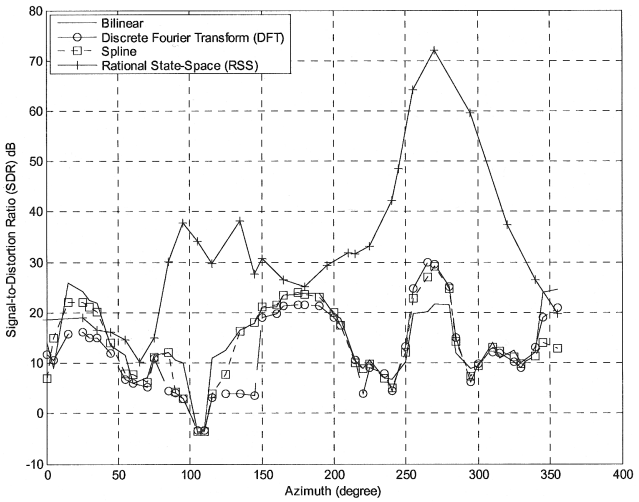

Figure 10. *Interpolation accuracy.*

4 Conclusion

We have addressed the binaural sound source localization problem at the teleoperator site, and the 3D sound synthesis at the operator site, both being integral parts within a general acoustic telepresence setting. We propose an efficient sound source localization method that demonstrates the ability of precise azimuth and elevation estimation, using a generic HRTF database. The

HRTF dataset is measured on the horizontal plane from 0° to 180° with 5° increments and on the vertical plane from −40° to 90° with 10° increments. The results indicate that, in a real environment, we can localize the sound source exactly at its target location 89% of the time. The remaining 11% that are falsely localized are within 10° on average from the target sound source location. If we construct the HRTF dataset with smaller increments, the resolution of estimation will be increased. The efficiency of the new algorithm suggests a cost-effective implementation for robot platforms and allows for a fast localization of single sound sources.

Furthermore, we have proposed a method for the interpolation of binaural impulse-responses based on the solution of a rational minimal state-space interpolation problem. Compared with existing interpolation techniques, this method allows very precise reconstruction of HRTFs in the horizontal plane and proved to have superior performance for a wide range of azimuths.

Based on the presented new method for sound source localization, future work is anticipated to include range estimation, sound separation, and classification. Moreover, due to the simplicity of the proposed localization algorithm, the integration of audio with other modalities, for example, haptic and vision, becomes promising. This kind of integration will allow the multi-sensory telepresence environment to improve the perceived degree of immersion for telepresence systems.

Acknowledgments

This work is fully supported by the German Research Foundation (DFG) within the collaborative research center SFB453 "High-Fidelity Telepresence and Teleaction," project area M4 "Acoustic Telepresence—Binaural Directional Hearing and Immersive Audio."

References

Ajdler, L. S. T., & Vetterli, M. (2005). The plenacoustic function on the circle with application to HRTF interpolation. *Proceedings of IEEE ICASSP, 3,* 273–276.

Altinsoy, E. (2003). Perceptual aspects of auditory-tactile asynchrony. *Proceedings of the Tenth International Congress on Sound and Vibration,* 3831–3838.

Anderson, B., & Antoulas, A. (1990). Rational interpolation and state-variable realizations. *Linear Algebra and Its Applications, 137,* 479–509.

Antoulas, A., & Anderson, B. (1986). On the scalar rational interpolation problem. *IMA Journal Binaural Source Localization and Spatial Audio for Telepresence Applications, 3,* 61–81.

Avanzini, F., Rocchesso, D., & Serafin, S. (2004). Friction sounds for sensory substitution. *Proceedings of the International Conference on Auditory Display, 4,* 1–8.

Begault, D. (2004). *3-D sound for virtual reality and multimedia*. San Diego, CA: Academic Press.

Beliczynski, B., Kale, I., & Cain, G. D. (1992). Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction. *IEEE Transaction Signal on Processing, 40*(3), 532–542.

Blauert, J. (1997). Spatial hearing: The psychophysics of human sound localization (rev. ed.). Cambridge, MA: MIT Press.

Duraiswami, R., Zhiyun, L., Zotkin, D. N., Grassi, E., & Gumerov, N. A. (2005). Plane-wave decomposition analysis for spherical microphone arrays. *IEEE WASPAA,* 150–153.

Gardner, W. G., & Martin, K. D. (1995). HRTF measurements of a KEMAR. *Journal of the Acoustical Society of America, 97*(6), 3907–3908.

Grassi, E., & Shamma, S. A. (2001). A biologically inspired, learning, sound localization algorithm. *Conference on Information Sciences and Systems,* 344–348.

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

Handzel, A. A. (2005). High acuity sound-source localization by means of a triangular spherical array. *Proceedings of IEEE ICASSP, 4,* 1057–1060.

Handzel, A. A., & Krishnaprasad, P. S. (2002). Biomimetic sound-source localization. *IEEE Sensors Journal, 2*(6), 607–616.

Horiuchi, T. K. (2005). "Seeing" in the dark: Neuromorphic VLSI modeling of bat echolocation. *IEEE Signal Processing Magazine,* 134–139.

Jahromi, O., & Aarabi, P. (2005). Theory and design of multirate sensor arrays. *IEEE Transactions on Signal Processing, 53*(5), 1739–1753.

Keyrouz, F., Naous, Y., & Diepold, K. (2006). A new method for binaural 3D localization based on HRTFs. *IEEE ICASSP.*

Kistler, D. J., & Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustics Society of America, 91*(3), 1637–1647.

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

Matsumoto, M., Yamanaka, S., & Tohyama, M. (2004). Effect of arrival time correction on the accuracy of binaural impulse response interpolation. *Journal of the Audio Engineering Society, 52,* 56–61.

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

Møller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics, 36*(3–4), 171–218.

Murray, J. C., Erwin, H., & Wermter, S. A. (2005). Recurrent neural network for sound-source motion tracking and prediction. *IEEE IJCNN, 4,* 2232–2236.

Nakashima, H., & Mukai, T. (2005). 3D Sound source localization system based on learning of binaural hearing. *IEEE International Conference on Systems, Man and Cybernetics, 4,* 3534–3539.

Robeson, S. (1999). Spherical methods for spatial interpolation: Review and evaluation. *Cartography and Geographic Information Systems, 24*(1), 3–20.

Rui, Y., & Florencio, D. (2003). New direct approaches to robust sound source localization. *Proceedings of IEEE ICME, 1,* 737–740.

Wang, Q. H., Ivanov, T., & Aarabi, P. (2004). Acoustic robot navigation using distributed microphone arrays. *Information Fusion (Special Issue on Robust Speech Processing), 5*(2), 131–140.