Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Sound and Vibration 281 (2005) 1093–1115
JOURNAL OF SOUND AND VIBRATION
www.elsevier.com/locate/jsvi

Head-related transfer function (HRTF) synthesis based on a three-dimensional array model and singular value decomposition

Mingsian R. Bai*, Kwuen-Yieng Ou

Department of Mechanical Engineering, National Chiao-Tung University, 1001 Ta-Hsueh Road, Hsin-Chu 300, Taiwan, Republic of China

Received 17 March 2003; received in revised form 21 October 2003; accepted 12 February 2004
Available online 29 September 2004

Abstract

An external ear model based on a three-dimensional array beamformer is presented to synthesize the head-related transfer functions (HRTFs). The array coefficients are calculated by matching the measured HRTFs with a frequency-domain template. In light of the characteristics of human hearing, a non-uniform sampling technique is also devised to reduce the problem to a manageable size. The model-matching problem is then solved, by using singular value decomposition procedure. Numerical simulations are carried out to investigate various array configurations in relation to approximation performance. A subjective experiment of sound localization is undertaken to demonstrate the validity of the proposed technique. From the results, the proposed beamformer-based technique proves to be an effective method for modeling the HRTFs.

1. Introduction

Three-dimensional (3D) sound reproduction has been a subject that received considerable attention in the audio signal-processing community. A common goal in this area is to create a more spatial and immersive audio impression for listeners, via either headphones or loudspeakers.

*Corresponding author. Fax: +886-3-5720634.
E-mail address: msbai@mail.nctu.edu.tw (M.R. Bai).

0022-460X/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsv.2004.02.007

3D audio reproduction has many applications such as virtual reality, robotics, teleconferencing, video-conferencing, entertainment, training simulation, and so forth.

A well-known duplex theory, due to Lord Rayleigh, suggests that humans' ability in sound localization relies heavily on the interaural time difference (ITD), and the interaural intensity difference (IID). ITD and IID refer to the time and level differences, respectively, between the sound signals received by the left and right ears. Apart from this simple model, more sophisticated models have been suggested to account for various shadowing and diffraction effects due to pinna, head, and torso. The head-related transfer function (HRTF) is such a model that treats these combined effects as a spatial filter that responds to the incoming sound field with different gain and phase in different frequency and incident angle [1–4]. HRTFs are typically documented as frequency response functions. Alternately, it can be documented as impulse response functions, or the head-related impulse response (HRIR). Hammershøi summarized the technical considerations on the acoustical and electro-acoustical transfer functions and calibration of the transmission chain during recording and reproduction of binaural signals [5], No matter which way is used, tremendous memory storage is necessary for the vast amount of data for all source directions in a 3D space. It is then highly desirable to develop a reduced-order model appropriate for real-time processing, without too much a performance penalty on human localization. In this regard, much research has been devoted to the development of reduced order HRTF models. Batteau suggested that the external ear could be modeled as a three-channels two-delay and sum acoustic coupler [6]. One delay unit varies with the source elevation, whereas the others vary with the source azimuth. Mackenzie et al. proposed a model based on tenth-order infinite impulse response (IIR) filters, obtained using balanced model truncation [7]. Genuit developed a filter-bank model that has 16 delay channels, based on the external ear geometry [8]. Duda also suggested structural models for HRTFs based on the spherical-head model and a monaural pinna model [9]. Kahana and Nelson investigated the basic far field radiation patterns of a baffled pinna using the singular value decomposition (SVD) and the boundary element method (BEM) [10]. Chen et al. proposed a functional representation for complex-valued HRTFs [11]. In this approach, HRTFs are represented by a weighted sum of spatial characteristic functions for all directions. In addition, Chen et al. also derived a broadband beamformer model from measured HRTFs [12]. In their method, the beamformer weights are chosen according to the characteristics of the external ears. The sensor geometry is essentially a 2D structure that accounts for only the variations on the azimuths of HRTFs. The beamformer approach has the advantages in that it enables substantial data compression and smooth interpolation for discrete directions, which is attractive for real-time audio rendering. However, the numberof weights increases rapidly according to the required modeling accuracy. This results in immense computational burden and numerical instability. It follows that only the azimuth angles have been simulated in Chen's work. Alternatively, low dimension and orthogonal representation for HRTFs measurement have been developed by applying the principal component analysis (PCA) to the logarithms of the HRTFs' magnitudes after the direction-independent, frequency dependency is removed [13,14].

Parallel to Chen et al.'s work, a modeling technique of HRTFs also based on the array beamformer idea is presented in this paper. However, there are several distinct features to the method presented herein to overcome the shortcomings mentioned before. First, the 2D array model is extended to a three-armed array for capturing the variations on the azimuths as well as elevations in a 3D space. Second, the array coefficients are calculated by matching the array

response with that of measured HRTFs non-uniformly sampled in the frequency domain. The frequency samples are determined by exploiting the characteristics of human hearing that resemble a constant-Q filter bank [15]. This non-uniform sampling method effectively alleviates the computation loading and numerical instability, as compared to the uniform sampling method [16]. Third, SVD in conjunction with a regularization procedure is proposed for solving the aforementioned model-matching problem. This method is also compared to a zeroth-order regularization procedure suggested in Ref. [17]. Numerical simulations and subjective experiments are carried out, using a HRTF database measured using a KEMAR by Gardner and Martin of the MIT Media laboratory [4], to demonstrate the validity of the proposed beamformer technique.

2. The 3D array model

Assume that the signal $r(t)$ received at a reference point is narrowband with center frequency ω_c:

$$r(t) = \mathrm{Re}\{s(t)\mathrm{e}^{\mathrm{j}\omega_c t}\}, \tag{1}$$

where $s(t)$ is the phasor of $r(t)$. Let $x_i(t)$ be the signal received at the ith array element located at $\vec{x}_i$, and $\vec{r}$ be the unit vector pointing to the source direction as shown in Fig. 1. If the speed of sound is c, the signal $x_i(t)$ can be written as

$$x_i(t) = r\left(t + \frac{\vec{x}_i \cdot \vec{r}}{c}\right) = \mathrm{Re}\left\{s\left(t + \frac{\vec{x}_i \cdot \vec{r}}{c}\right)\mathrm{e}^{\mathrm{j}\omega_c \vec{x}_i \cdot \vec{r}/c}\mathrm{e}^{\mathrm{j}\omega_c t}\right\}. \tag{2}$$

In general, $s(t + \vec{x}_i \cdot \vec{r}/c) \approx s(t)$ for far-field approximation. For M sensor signals $x_1(t), \ldots, x_M(t)$, the data vector can be formed as

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} \mathrm{e}^{\mathrm{j}k_c \vec{x}_1 \cdot \vec{r}} \\ \vdots \\ \mathrm{e}^{\mathrm{j}k_c \vec{x}_M \cdot \vec{r}} \end{bmatrix} s(t)\mathrm{e}^{\mathrm{j}\omega_c t} = \mathbf{a}(\vec{r})r(t), \tag{3}$$

where $k_c = \omega_c/c = 2\pi/\lambda_c$ is the wavenumber, with λ_c being the wavelength. The vector $\mathbf{a}(\vec{r})$ is called the array manifold vector.

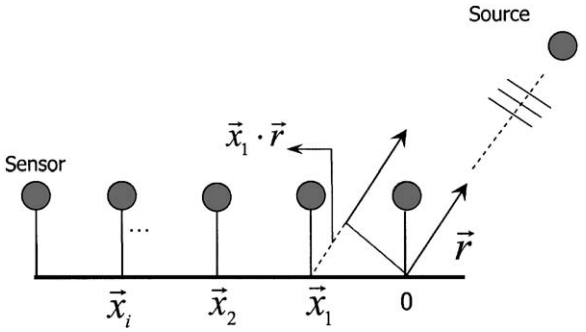


Fig. 1. The schematic diagram of a uniformly linear array.

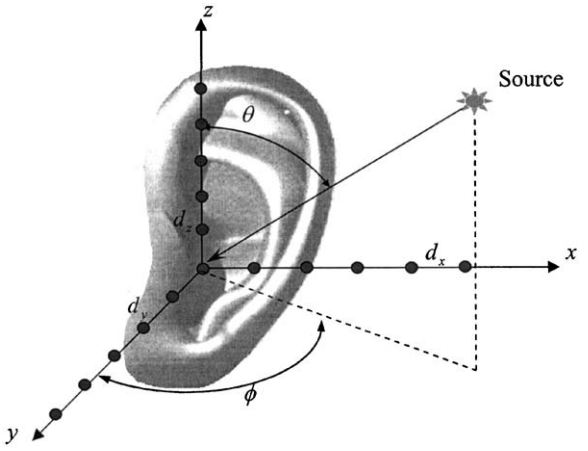


Fig. 2. The array geometry of the 3D external ear model.

The L-shaped array proposed by Chen et al. is able to model the HRTFs for azimuths varying from 0° to 90° [12]. This paper attempts to extend this beamformer concept by devising a three-armed array to account for both the azimuths and the elevations, such that the characteristics of the external ear can be better addressed. The geometry of the proposed array is shown in Fig. 2. The unit vector $\vec{r}$ for a source at the (θ, ϕ) direction is given by

$$\vec{r} = (\sin\theta \sin\phi, \sin\theta \cos\phi, \cos\theta). \tag{4}$$

Assuming uniform spacing d_x, d_y, and d_z for each arm, the position vector of the ith element can be expressed as

$$\vec{x}_i = ((L_x - 1)d_x, (L_y - 1)d_y, (L_z - 1)d_z), \tag{5}$$

where (L_x, L_y, L_z) is the element index along the x, y, and z axes, respectively. Then,

$$\vec{x}_i \cdot \vec{r} = (L_x - 1)d_x \sin\theta \sin\phi + (L_y - 1)d_y \sin\theta \cos\phi + (L_z - 1)d_z \cos\theta. \tag{6}$$

Assume that each arm has M sensor elements, plus one element at the origin, the array manifold vector can be written as a $(3M + 1) \times 1$ vector

$$\begin{aligned}\mathbf{a}_n(\omega_c, \theta, \phi) = [1,\ & e^{jk_c d_x \sin\theta \sin\phi}, \ldots, e^{jk_c M d_x \sin\theta \sin\phi}, e^{jk_c d_y \sin\theta \cos\phi}, \ldots, \\ & e^{jk_c M d_y \sin\theta \cos\phi}, e^{jk_c d_z \cos\theta}, \ldots, e^{jk_c M d_z \cos\theta}]^{\mathrm{T}}.\end{aligned} \tag{7}$$

Extension from the narrowband formulation to the broadband formulation is straightforward. With reference to Fig. 3, the beamformer output $y(t)$ is the weighted sum of the delayed input signals $x_i(t)$, $i = 1, \ldots, 3M + 1$, i.e.

$$y(t) = \sum_{m=1}^{3M+1} \sum_{n=0}^{N} w_{mn}^* x_m(t - nT) = \mathbf{w}^{\mathrm{H}} \mathbf{x}(t), \tag{8}$$

where $\mathbf{w} = [w_{10}, \ldots, w_{1N}, \ldots, w_{(3M+1)0}, \ldots, w_{(3M+1)N}]^{\mathrm{H}}$ and $\mathbf{x}(t) = [x_1(t), \ldots, x_1(t - NT), x_2(t), \ldots, x_2(t - NT), \ldots, x_{3M+1}(t), \ldots, x_{3M+1}(t - NT)]^{\mathrm{T}}$. The size of $\mathbf{w}$ is $[(N + 1)(3M + 1)] \times 1$. Eq. (8) can

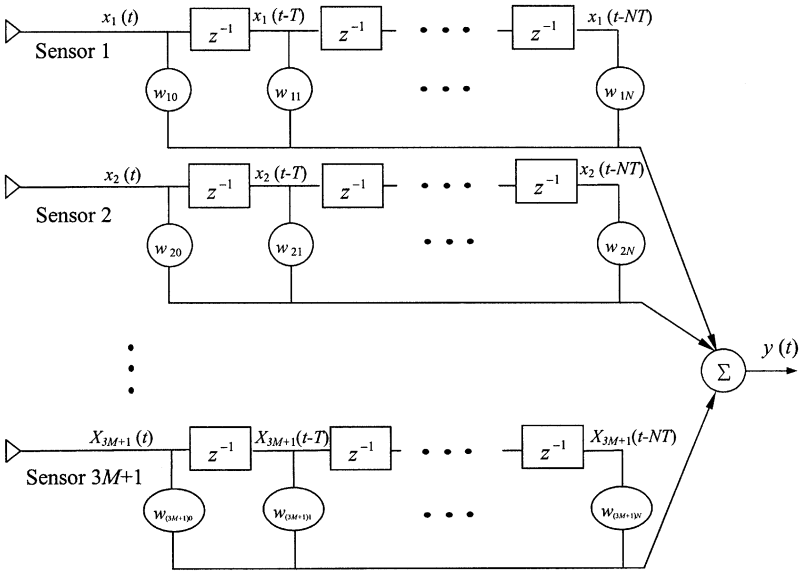


Fig. 3. The block diagram of the broadband beamformer model. FIR filters are incorporated in each sensor channel.

be rewritten in the frequency domain for a particular direction (θ, ϕ) as

$$y(\omega_c, \theta, \phi) = \mathbf{w}^{\mathrm{H}}\mathbf{x}(\omega_c, \theta, \phi) = \mathbf{w}^{\mathrm{H}}\mathbf{a}(\omega_c, \theta, \phi)r(\omega_c), \tag{9}$$

where the manifold vector is given by

$$\begin{aligned}\mathbf{a}(\omega_c, \theta, \phi) = [&\mathbf{F}(\omega_c),\ \mathbf{F}(\omega_c)\mathrm{e}^{\mathrm{j}k_c d_x \sin\theta\sin\phi}, \ldots, \mathbf{F}(\omega_c)\mathrm{e}^{\mathrm{j}k_c M d_x \sin\theta\sin\phi},\ \mathbf{F}(\omega_c)\mathrm{e}^{\mathrm{j}k_c d_y \sin\theta\cos\phi}, \ldots,\\ &\mathbf{F}(\omega_c)\mathrm{e}^{\mathrm{j}k_c M d_y \sin\theta\cos\phi},\ \mathbf{F}(\omega_c)\mathrm{e}^{\mathrm{j}k_c d_z \cos\theta}, \ldots, \mathbf{F}(\omega_c)\mathrm{e}^{\mathrm{j}k_c M d_z \cos\theta}]^{\mathrm{T}}\end{aligned} \tag{10}$$

with $\mathbf{F}(\omega_c) = [1,\ \mathrm{e}^{\mathrm{j}\omega_c T}, \ldots, \mathrm{e}^{\mathrm{j}\omega_c NT}]$ and N and T being the filter order and the sampling period, respectively. The dimensions of $\mathbf{F}(\omega_c)$ and $\mathbf{a}(\omega_c, \theta, \phi)$ are $1 \times (N+1)$ and $[(N+1)(3M+1)] \times 1$, respectively. It can be observed from Eq. (9) that $\mathbf{w}^{\mathrm{H}}\mathbf{a}(\omega_c, \theta, \phi)$ accounts for transfer relation between the source r and the array output y.

In our problem, we seek to find an array model that closely matches the radiation pattern of the measured HRTFs. This is accomplished by solving the following optimization problem

$$\min_{\mathbf{w}} \|h_d(\omega_c, \theta, \phi) - \mathbf{w}^{\mathrm{H}}\mathbf{a}(\omega_c, \theta, \phi)\| \tag{11}$$

where $h_d(\omega_c, \theta, \phi)$ is the desired HRTF to be matched by the array model. The minimum value of Eq. (11) can be achieved if $\mathbf{w}^{\mathrm{H}}\mathbf{a}(\omega_c, \theta, \phi) \approx h_d(\omega_c, \theta, \phi)$. By discretization in the frequency and angle domains, i.e., ω_i, $i = 1, 2, \ldots, P$, θ_j, $j = 1, 2, \ldots, Q$ and ϕ_k, $k = 1, 2, \ldots, R$, Eq. (11) can be rewritten into the following matrix equation:

$$\mathbf{D}^{\mathrm{H}}\mathbf{w} \approx \mathbf{h}_d, \tag{12}$$

where

$$\mathbf{h}_d = [h_d(\omega_1, \theta_1, \phi_1), \ldots, h_d(\omega_p, \theta_1, \phi_1), \ldots, h_d(\omega_p, \theta_Q, \phi_R)]^{\mathrm{H}},$$

$$\mathbf{D} = [\mathbf{A}(\theta_1, \phi_1), \ldots, \mathbf{A}(\theta_Q, \phi_1),\ \mathbf{A}(\theta_1, \phi_2), \ldots, \mathbf{A}(\theta_Q, \phi_2), \ldots, \mathbf{A}(\theta_Q, \phi_R)],$$

with

$$\mathbf{A}(\theta_j, \phi_k) = [\mathbf{a}(\omega_1, \theta_j, \phi_k),\ \ \mathbf{a}(\omega_2, \theta_j, \phi_k), \ldots, \mathbf{a}(\omega_p, \theta_j, \phi_k)].$$

The sizes of $\mathbf{A}(\theta_j, \phi_k)$, $\mathbf{D}$ and $\mathbf{h}_d$ are $[(N+1)(3M+1)] \times P$, $[(N+1)(3M+1)] \times (PQR)$ and $PQR \times 1$, respectively.

3. Optimization of array design

3.1. Model matching by the least-square approach

In this section, the model-matching problem for the array design, described by the linear system of Eq. (12), is solved by using a least-square procedure. To ensure that the array coefficient $\mathbf{w}$ is real [12], we split the complex equation (12) into real and imaginary parts:

$$\mathbf{D}_R^{\mathrm{T}}\mathbf{w} = \mathbf{h}_{dR} \tag{13}$$

and

$$\mathbf{D}_I^{\mathrm{T}}\mathbf{w} = \mathbf{h}_{dI}, \tag{14}$$

where the subscripts "R" and "I" denote the real and imaginary parts, respectively. Eqs. (13) and (14) can be further assembled into a single matrix equation

$$\mathbf{C}\mathbf{w} = \mathbf{g}, \tag{15}$$

where $\mathbf{g} = [\mathbf{h}_{dR}^{\mathrm{T}}, \mathbf{h}_{dI}^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{C} = [\mathbf{D}_R, \mathbf{D}_I]^{\mathrm{T}}$. The sizes of $\mathbf{g}$ and $\mathbf{C}$ are $(2PQR) \times 1$ and $(2PQR) \times [(N+1)(3M+1)]$, respectively. However, the matrix $\mathbf{C}$ is generally not Hermitian, and even not square. The inverse of $\mathbf{C}$ will not exist in this case. Kirkeby et al. [17] proposed a regularization procedure which combines the least-squares inversion and the zeroth-order regularization to find the optimal solution of $\mathbf{w}$ in Eq. (15):

$$\mathbf{w} = (\mathbf{C}^{\mathrm{T}}\mathbf{C} + \beta\mathbf{I})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{g}. \tag{16}$$

The parameter β is used to control the degree of regularization. The exact value of β is usually not critical and thus is subject to judicious choice.

3.2. Model-matching by the SVD with regularization

In this paper, a modified regularization procedure based on the SVD is proposed. The SVD of the matrix $\mathbf{C}$ is [18]

$$\mathbf{C} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{H}}, \tag{17}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $(2PQR)\times(2PQR)$ and $[(N+1)(3M+1)]\times[(N+1)(3M+1)]$ unitary matrices, respectively. $\mathbf{\Sigma}$ is a $(2PQR)\times[(N+1)(3M+1)]$ matrix whose entries are all zeros except for the non-negative diagonal elements called the singular values. The number of the singular values, $\sigma_1>\sigma_2>\cdots>\sigma_\rho$, is the rank of the matrix $\mathbf{C}$, denoted by ρ and $\rho<(N+1)(3M+1)$. The least-square solution of $\mathbf{w}$ in Eq. (15) is written as

$$\mathbf{w}=\sum_{i=0}^{\rho}\sigma_i^{-1}\mathbf{v}_i\mathbf{u}_i^{\mathrm{T}}\mathbf{g}, \tag{18}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the left and right singular vectors. If $\mathbf{C}$ is rank deficient, the inverse of small singular values often results in excess magnitude of $\mathbf{w}$. Thus, it is common practice to retain only significant singular values, according to a prescribed threshold, in solving for $\mathbf{w}$.

A drawback inherent to the above-mentioned procedure can be explained by drawing an analogy between SVD and a frequency response function. The singular values and the unitary matrices can be related, respectively, to the magnitude and phase exponential of the frequency response function. Direct truncation of the singular values may produce large distortion in high-frequency range. As a remedy to this commonly used approach, a modified regularization method that replaces small singular values by a constant δ is proposed in this paper. The singular value less than a specified threshold δ is replaced by δ. δ can be chosen to be a small number, say, 0.1% of the maximal singular value, as a rule of thumb. With this modification, the solution of $\mathbf{w}$ is calculated

$$\mathbf{w}=\sum_{i=0}^{(N+1)(3M+1)}\sigma_i^{-1}\mathbf{v}_i\mathbf{u}_i^{\mathrm{T}}\mathbf{g}. \tag{19}$$

3.3. Non-uniform spectral sampling

It is well known that the human auditory system performs a non-uniform spectral analysis similar to a constant-Q filter bank on the received sound waves [15,16]. For the human hearing, excessively high spectral resolution in the high-frequency range is hardly necessary. In this paper, a non-uniform sampling technique is presented by exploiting this property of human hearing. The key step of this technique is to sample the frequency response function at exponentially spaced points, according to

$$\widetilde{\omega}_i=\omega_0 e^{a(2\pi/P)i},\quad i=1,\ldots,P, \tag{20}$$

where ω_0 and P are the lowest frequency and desired number of frequency components, respectively. The parameter a must be chosen to cover the desired frequency span $\omega_{\max}$,

$$a=\frac{1}{2\pi}\ln\left(\frac{\omega_{\max}}{\omega_0}\right). \tag{21}$$

After the model matching and the array coefficient $\mathbf{w}$ is determined, the frequency response of the HRTF in any direction (θ_j,ϕ_k) can be simulated by

$$\mathbf{h}(\theta_j,\phi_k)=\mathbf{w}^{\mathrm{T}}\mathbf{A}(\theta_j,\phi_k), \tag{22}$$

where $\mathbf{h}(\theta_j, \phi_k) = [h_d(\omega_1, \theta_j, \phi_k), \ldots, h_d(\omega_p, \theta_j, \phi_k)]$ is the frequency response vector in that direction, and $\mathbf{A}(\theta_j, \phi_k)$ has been defined earlier in the broadband beamformer. The HRTF in any direction is finally recovered, followed by adding the corresponding ITD for that direction. Using the array model, only several hundred weights and the ITD data for each direction need to be stored. In contrast to the direct implementation of the HRTF, where all directions, e.g., 710, with 256 taps each need to be maintained in a database, the present array approach provides an distinct advantage in terms of memory storage.

3.4. Optimization of array configuration

In this section, optimization techniques are employed to find the optimal array configuration. In particular, effects of parameters such as the number of weights, the number of sensors and filter taps are examined through the simulations. Since the model-matching problem in this paper is formulated with the least-square approach and the resulting cost function is quadratic, the global minimum of error is guaranteed. In the first and second simulations, the HRTF database of MIT includes azimuth data varying from 0° to 180° and elevation angle is set to zero. Each sensor was sampled uniformly in the frequency domain at 256 points over the bandwidth 22 kHz. In addition, the weight vector is obtained by the common SVD procedure [12]. To facilitate the comparison, the following normalized error is defined:

$$\mathrm{NER} = \frac{1}{QR}\sum_{k=1}^{R}\sum_{j=1}^{Q}\frac{\sum_{i=1}^{P}\left|h_d(\omega_i, \theta_j, \phi_k) - \mathbf{w}^{\mathrm{T}}\mathbf{a}(\omega_i, \theta_j, \phi_k)\right|^2}{\sum_{i=1}^{P}\left|h_d(\omega_i, \theta_j, \phi_k)\right|^2}. \tag{23}$$

In the first simulation, the number of array weights is fixed; the numbers of sensors and taps are varied in such a way that their product is a constant (300). This simulation is based on a linear array applied to the HRTFs on the horizontal plane. The normalized errors of seven configurations are shown in Fig. 4(a). The spacing between sensors is 0.8 mm. The results indicate that the configuration of 6-sensor in a line has the best performance.

In the second simulation, an L-shaped array similar to that used by Chen et al. [12] is used in this case because of its resemblance to the human pinna when viewed in the horizontal plane. The filter taps of each sensor are set to 50. The various ways of distributing the sensor along each arm of the array is examined while the total number of the sensors, including the one at the corner, is maintained to be 11. This simulation is also performed using the HRTFs on the horizontal plane. Fig. 4(b) illustrates the error versus the number of sensors on each axis of the L-shaped array. The result reveals that error is minimal when identical number of sensors, i.e., a 5-1-5 configuration, are used for each axis.

To better account for the variations of HRTFs in the elevation directions, a third arm perpendicular to those two arms in the L-shaped array is added to form a three-armed array. The number of sensors alone in each arm on the horizontal plane is held constant (5-1-5). The filter taps of each sensor are set to 50. This simulation is also performed using the HRTFs in the range, azimuth $= -30°$ and elevation $= -40$–$90°$. The plot of error versus number of sensors on the z-axis is shown in Fig. 5(a). Although the error decreases with the increasing number of sensors, no significant improvement seems to be obtained by using more than five sensors.

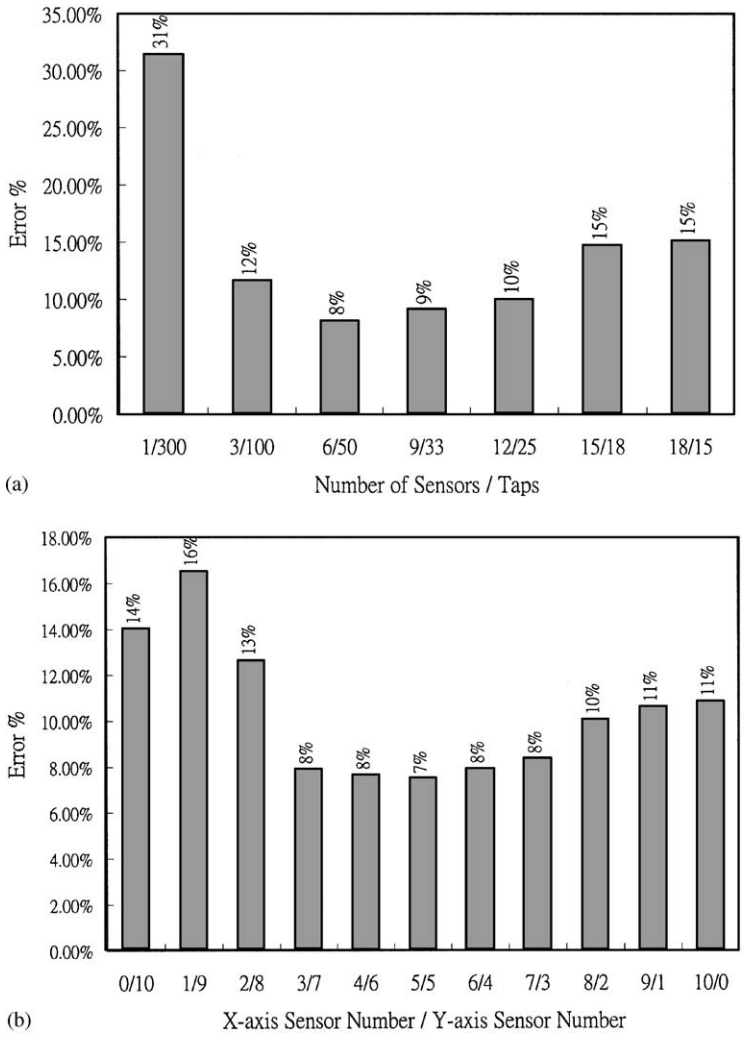


Fig. 4. Approximation error of the array model. (a) The effect of the number of sensors and the number of filter taps of a linear array. The number of the array coefficients is maintained fixed (300). (b) The effect of distributing sensors along each axis of an L-shaped array. The number of the total array sensors is maintained fixed (11).

With the observations in the above simulations, the array configuration is thus fixed to be a three-armed array with five sensors along each axis plus one at the origin. It is worth mentioning that this is a sensible engineering choice rather than an optimal one. The effect of filter length on the matching performance is next investigated. This simulation is performed on the basis of all 710 HRTFs of the MIT database. The result in Fig. 5(b) indicates that increasing the number of filter taps achieves almost no further improvement in matching performance when the filter is longer than 50 taps. Incidentally, this result seems to coincide with that of the first simulation. In what follows, we shall dwell on the 16-sensor array model with five sensors on each axis, plus one at the origin, which appears adequate to compromise between modeling error and complexity. The

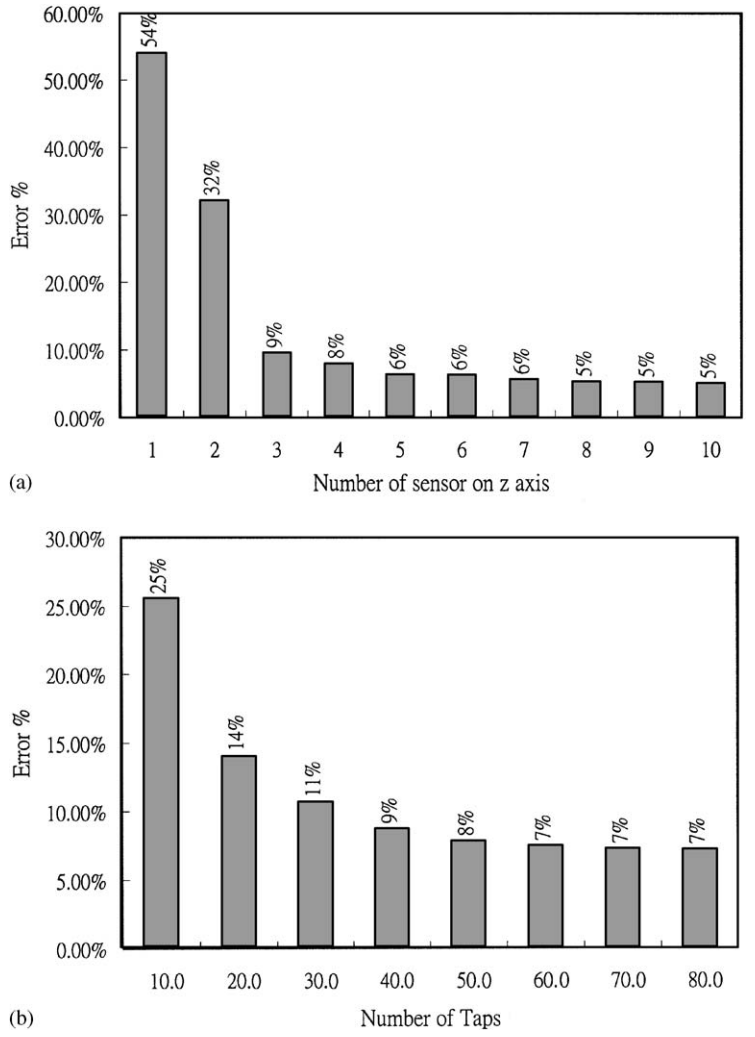


Fig. 5. Approximation error of the array model. (a) The effect of the number of sensors on the z-axis. Each arm of the x and y axes has five sensors, plus the one at the origin. (b) The effect of the number of filter taps for the 5-1-5 L-shaped array configuration.

spacing between sensors is 8 mm. Each sensor is connected to a filter with 50 taps. The complete array configuration is illustrated in Figs. 2 and 3.

4. Experimental verification

Experiments were carried out to evaluate the adequacy of the array beamformer model proposed in this paper. The HRTF database of MIT including azimuth data varying from 0° to

180° and elevation data varying form −30° to 90° was employed in the following tests. There is a subtle adjustment on the phase made to the database before constructing the array model. The effective delay between the stimulus arrival times at two ears has been obscured by the normalization process with the HRTF database. In this paper, the following technique based on the discrete Hilbert transform [19,20] is utilized to resolve the phase ambiguity of $h_d(\omega_i, \theta_j, \phi_k)$:

$$h_{\min}(i) = \exp\{\mathrm{FFT}[2\varepsilon(n)\mathrm{IFFT}\ \ln(|h(i)|)]\}, \tag{24}$$

where n and i are the discrete and frequency indices, respectively; ln denotes the natural logarithm and $\varepsilon(n)$ is given by $\varepsilon(n) = 0$ for $n<0$, $\varepsilon(n) = 1$ for $n>0$, and $\varepsilon(0) = 1/2$. FFT and IFFT symbolize the fast Fourier transform and the inverse fast Fourier transform, respectively. The term $h_{\min}(i)$ represents the ''minimum phase'' component of the discrete frequency response function of $h(i)$, or $h_d(\omega_i, \theta_j, \phi_k)$. The original HRTF is then replaced by its minimum-phase component cascaded with a linear phase shift due to the ITD. Fig. 6(a) and (b) illustrates the effect of this Hilbert transform procedure.

Subjective localization tests are conducted to verify the array model with 15 participants, using a white noise signal. The assessment criterion lies in the ability of localizing virtual sources using headphones. Fig. 7(a) and (b) depicts the geometrical arrangement of sources in relation to the listener in azimuth and elevation angles, respectively. The sound source is switched on for 2 s at each direction. Information regarding source locations and the switching pattern are made known to the participants before the test. Fig. 8 shows the results for the sound sources generated using the original HRTF in azimuth and elevation directions. The result appears quite normal except for some front–back reversals and larger localization error in elevations than in azimuths. Many researchers have carried out such tests. For example, the work done by Kahana et al. [21] has also been added to the references in this paper.

4.1. The least-square approach

The least-square method of Eq. (16) is employed to calculate the 16-sensored array model. The response of the 50-tapped filter connected to each sensor was sampled uniformly in the frequency domain at 256 points over a bandwidth of 22 kHz. In this case, the parameter β is set to be 8×10^{-5}. The filter coefficients are obtained via IFFT using uniform frequency samples. The matching error NER was found to be 12.47%. The localization test is then carried out to assess the performance of this array model. The result is shown in Fig. 9(a). As compared to the results of Fig. 8, the localization performance achieved by the least-square array is slightly worse than the original HRTF.

Next, a change is made to the least-square method in which the filter coefficients are obtained via IFFT using uniform frequency samples. In Eq. (20), the relevant parameters are: $\omega_0 = 2\pi \times 44$, $P = 256$, and $a = 0.993$. The response of the 50-tapped filter connected to each sensor was sampled non-uniformly in the frequency domain at 256 points over a bandwidth of 22 kHz. The matching error NER was found to be 12.03%. The results of the localization test are shown in Fig. 9(b). Comparison of Fig. 9(a), (b) and even Fig. 8 reveals that this non-uniform sampling technique enables to produce comparable performance as the original HRTF, and better performance than the uniformly sampled HRTF.

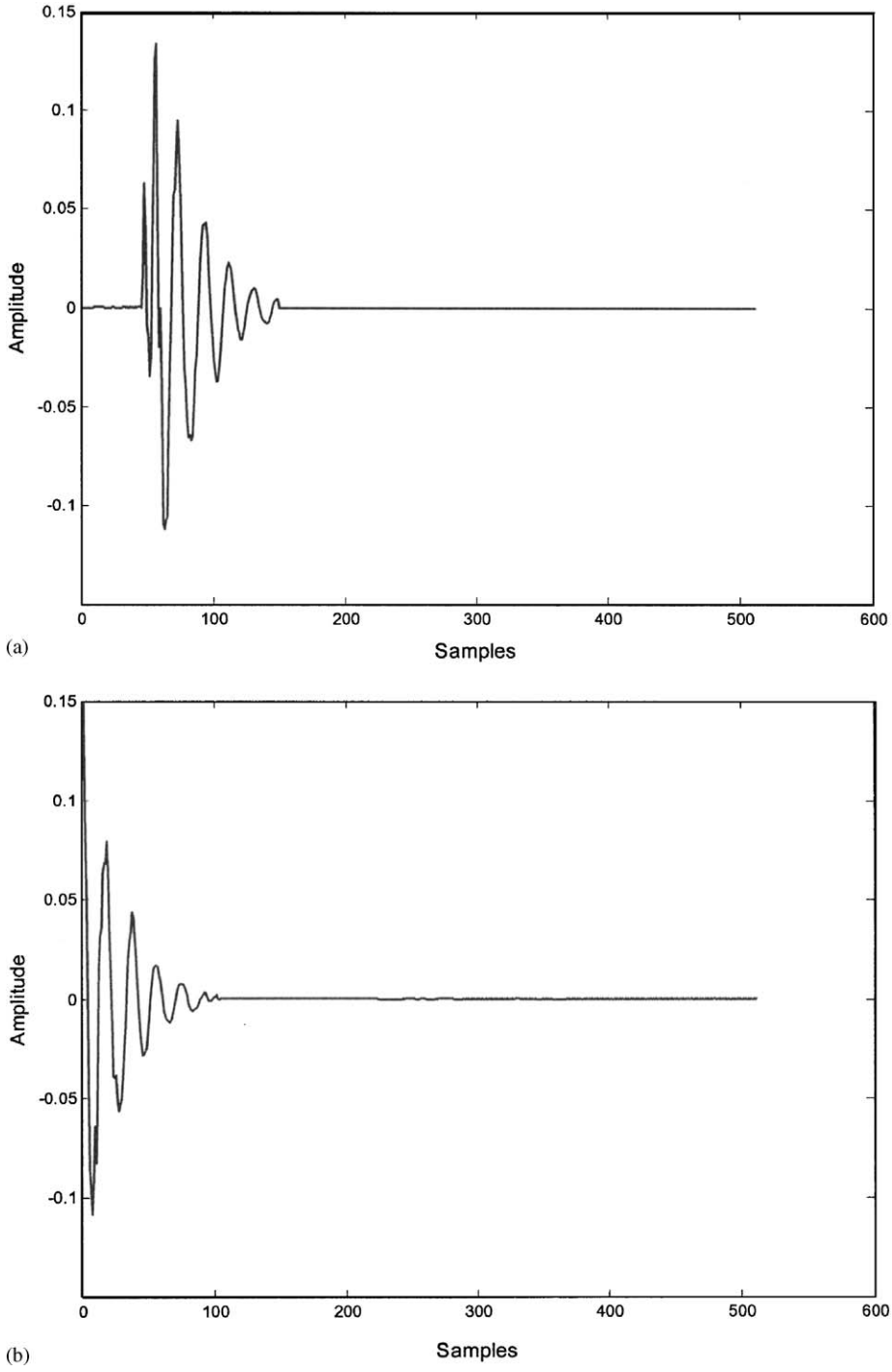


Fig. 6. The impulse response of a HRTF at azimuth 30° and elevation 0°. (a) The original HRTF. (b) Minimum-phase component obtained using the discrete Hilbert transform.

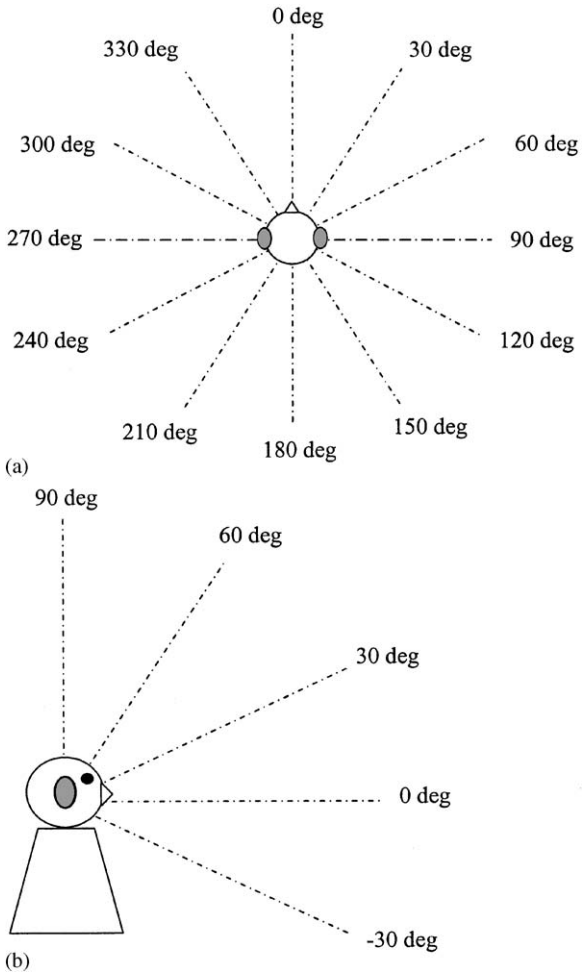


Fig. 7. The geometric arrangement of the subjective localization tests. (a) Horizontal plane. (b) Median plane.

4.2. The modified SVD procedure with regularization

The SVD method of Eq. (19) is used to calculate the 16-sensored array model. In this case, δ is set to be 0.1% of the maximum singular value. In model matching, the response of the 50-tapped filter connected to each sensor was sampled uniformly in the frequency domain at 256 points over a bandwidth of 22 kHz. The matching error NER was found to be 3.52%. In comparison to the 7.46% of the SVD with direct truncation and the 12.47% of the least-square approach, the modified SVD with regularization appears to provide some advantages approximating the HRTF as an array model. For simplicity, only the cases with the best matching performance are showed in Fig. 10(a) and (b). In addition, the non-uniform sampling method is also used for modeling the array. In this case, ω_0 is $2\pi \times 38$, P is 64, and a is set to be 1.0284 in Eq. (20). The matching error

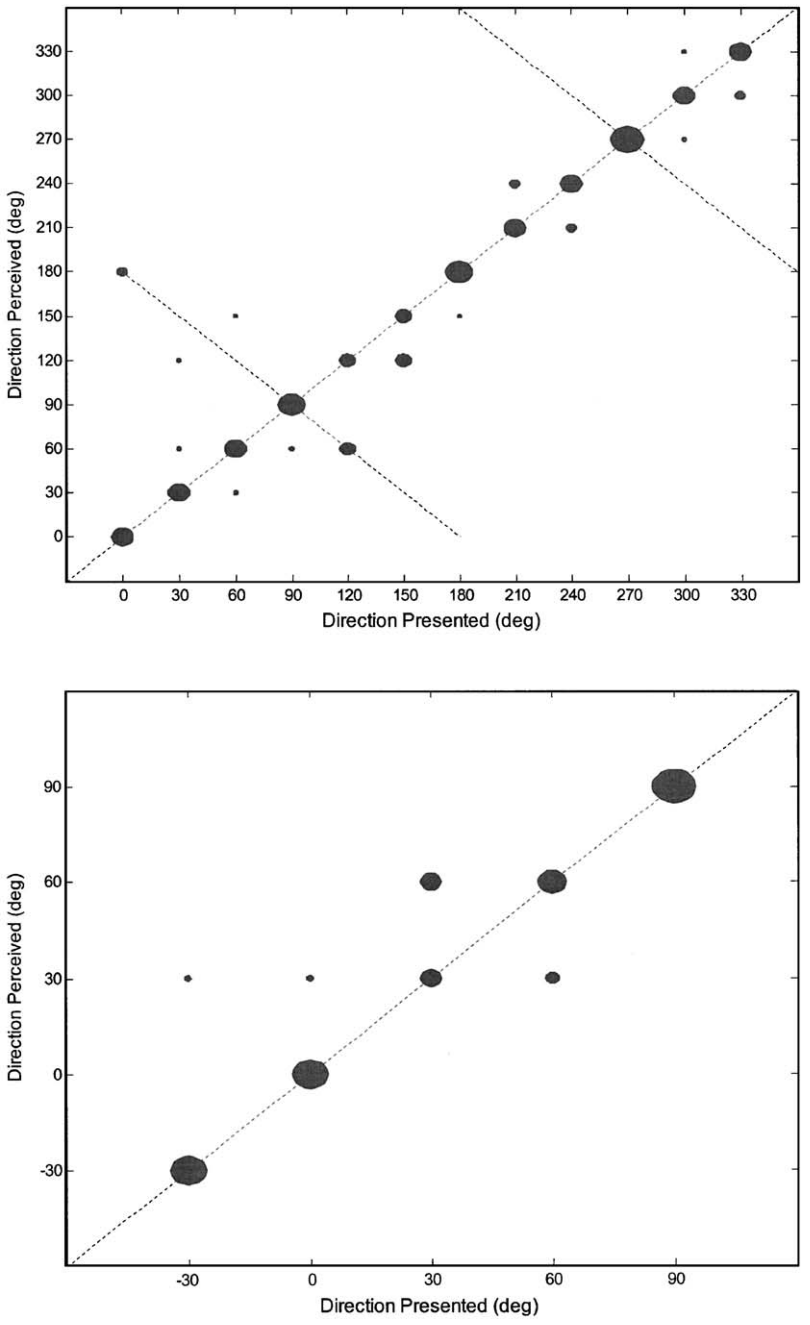


Fig. 8. The result of the subjective localization experiments for the original HRTFs. The abscissa is the direction where the source is presented and the ordinate is the direction perceived by the listener. The radius of each circle is proportional to the number of times of the direction perceived by the listeners.

NER was found to be 3.32%. Fig. 11(a) shows the subjective localization results in azimuth and elevation directions obtained using the modified SVD regularization with uniform sampling. Compared to Fig. 8, the performance is slightly worse than the original HRTF. Fig. 11(b) shows

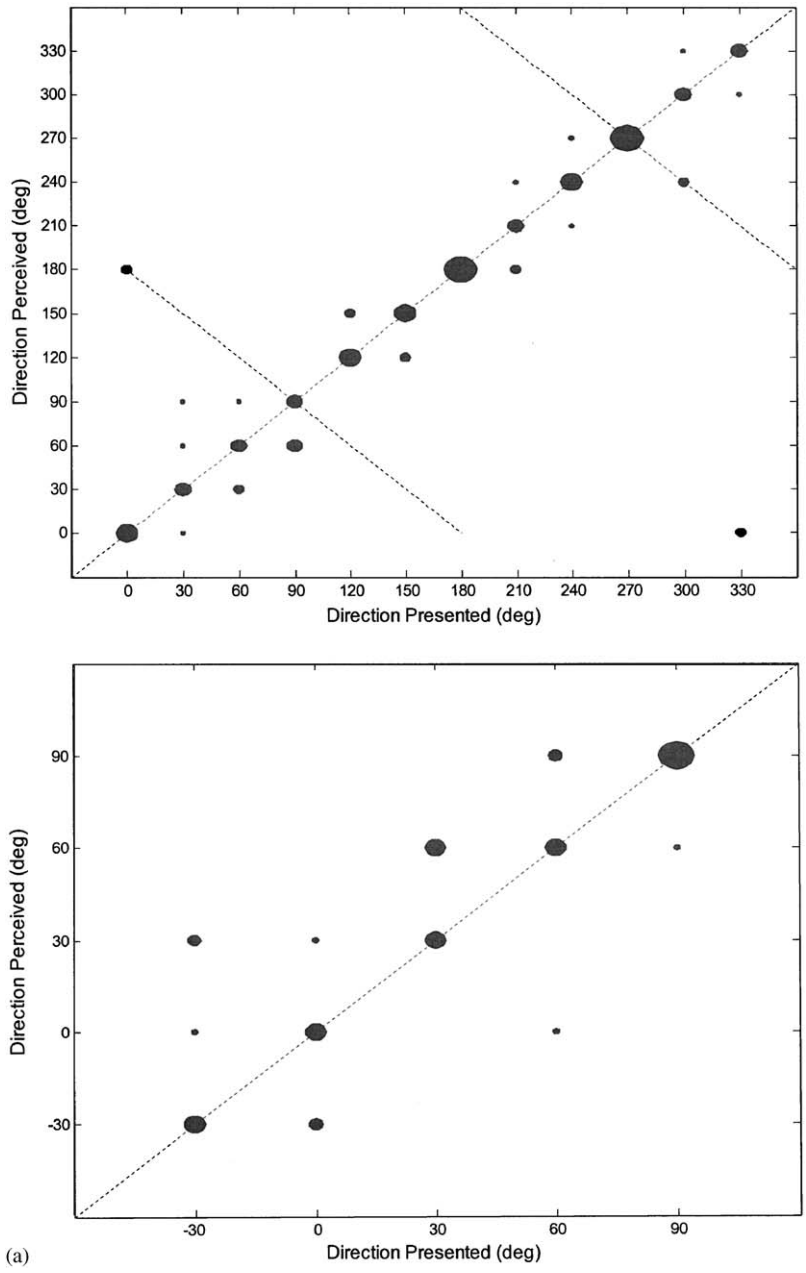


Fig. 9. The result of the subjective localization experiments for the least-square approach. (a) Uniform sampling approach. (b) Non-uniform sampling approach.

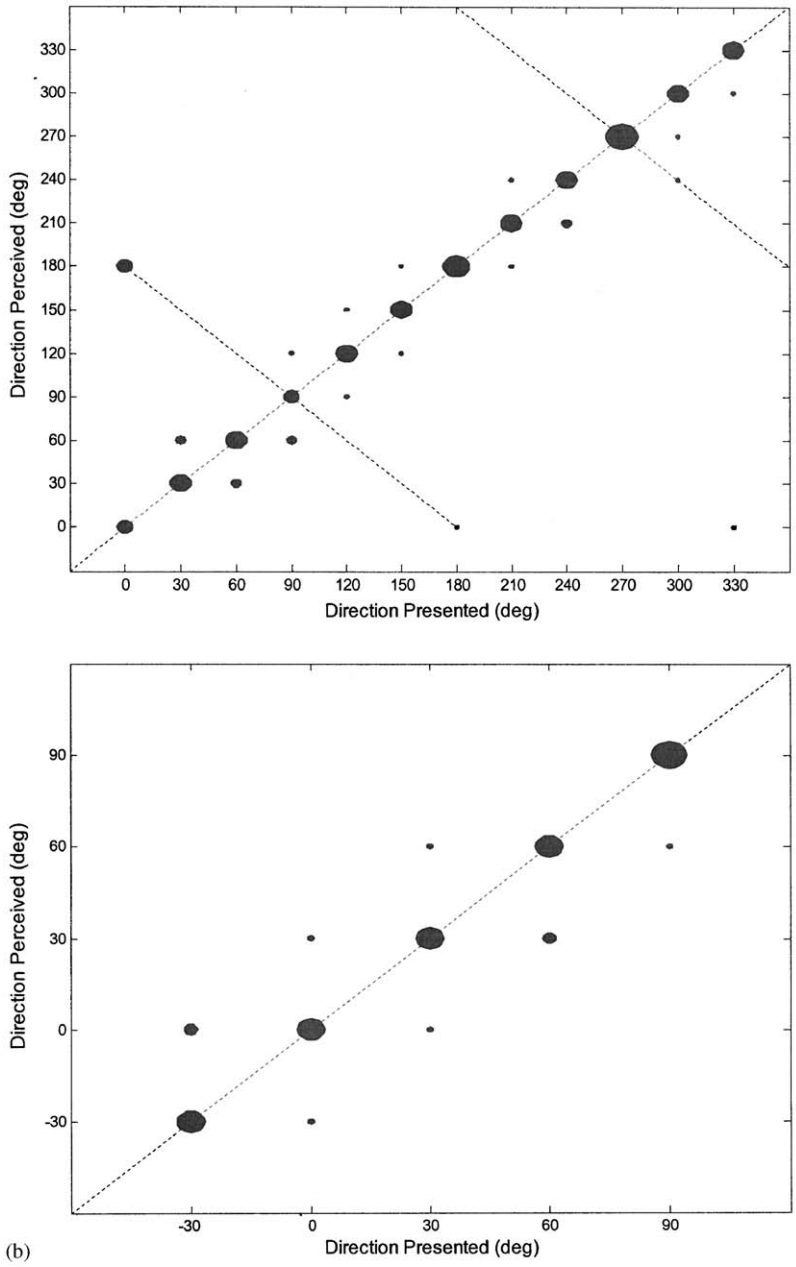


Fig. 9. (*Continued*)

the results obtained using the non-uniform sampling method. Again, the non-uniform sampling technique exhibits comparable performance as the original HRTF, and better performance than the uniformly sampled HRTF. An interesting comparison was also made between the present beamformer approach and the spherical-head model [9]. The results of a listening test obtained using Duda's method are shown in Fig. 12. From the results, it found that spherical-head model

suffers severer front–back reversals and larger localization error than the proposed methods. To better identify the differences among the HRTF implementations in the subjective tests, quantitative measures of the error mean (μ) and standard deviation (σ) for 15 subjects and 17 angles have been calculated and summarized in Table 1:

$$\mu = \frac{1}{15 \times 17} \sum_{i=1}^{15} \sum_{j=1}^{17} e_{ij}, \tag{25}$$

$$\sigma = \sqrt{\frac{1}{15 \times 17} \sum_{i=1}^{15} \sum_{j=1}^{17} (e_{ij} - \mu)^2}, \tag{26}$$

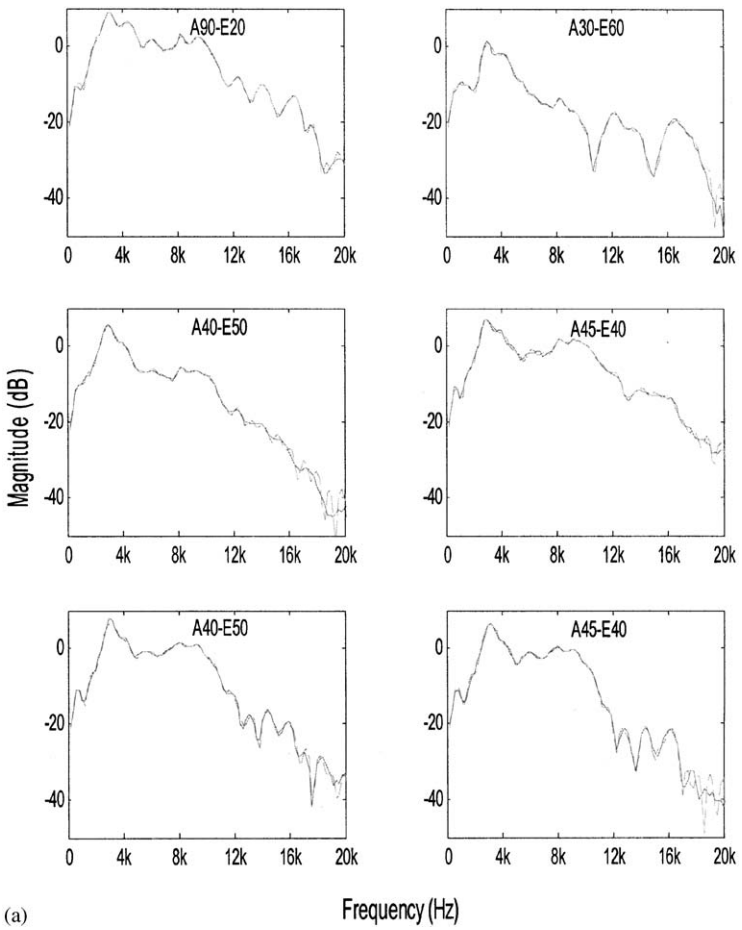


Fig. 10. Comparison of the desired (solid line) and the modeled (dashed line) magnitude and phase responses in six directions, obtained using the modified SVD method with regularization. The corresponding angle is marked in the figure: "A" and "E" represent the azimuth and the elevation, respectively. For example, A0-E30 signifies that azimuth $= 0^\circ$ and elevation $= 30^\circ$. (a) Magnitude responses scaled within -50 to 10 dB in the frequency range 0 to 20 kHz. (b) Phase responses scaled within -5 to 2 rad in the frequency range 0–20 kHz.

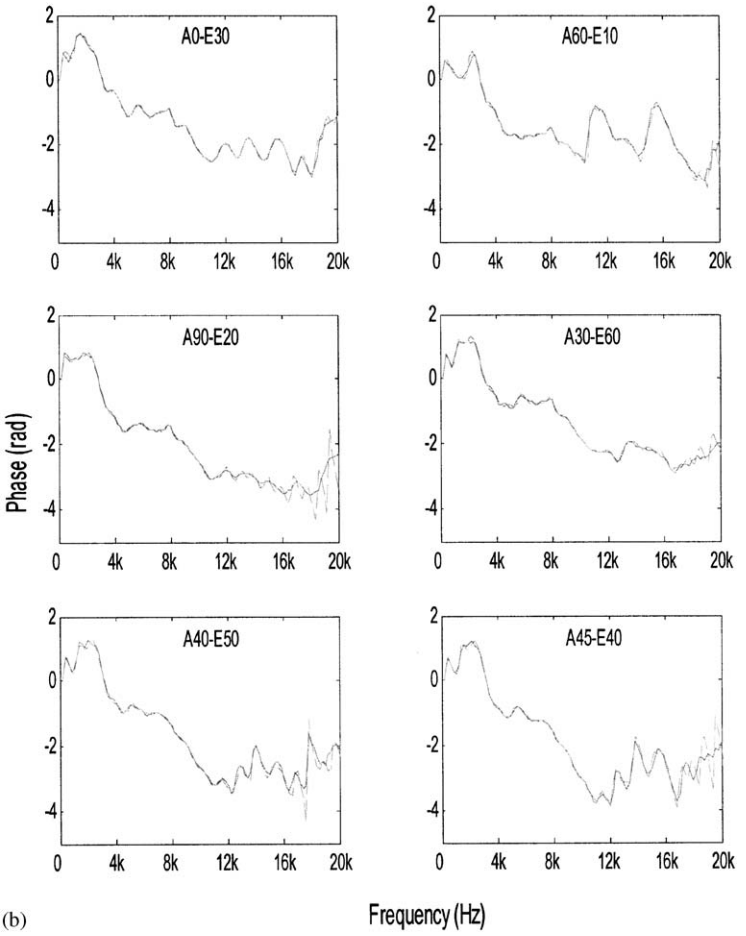
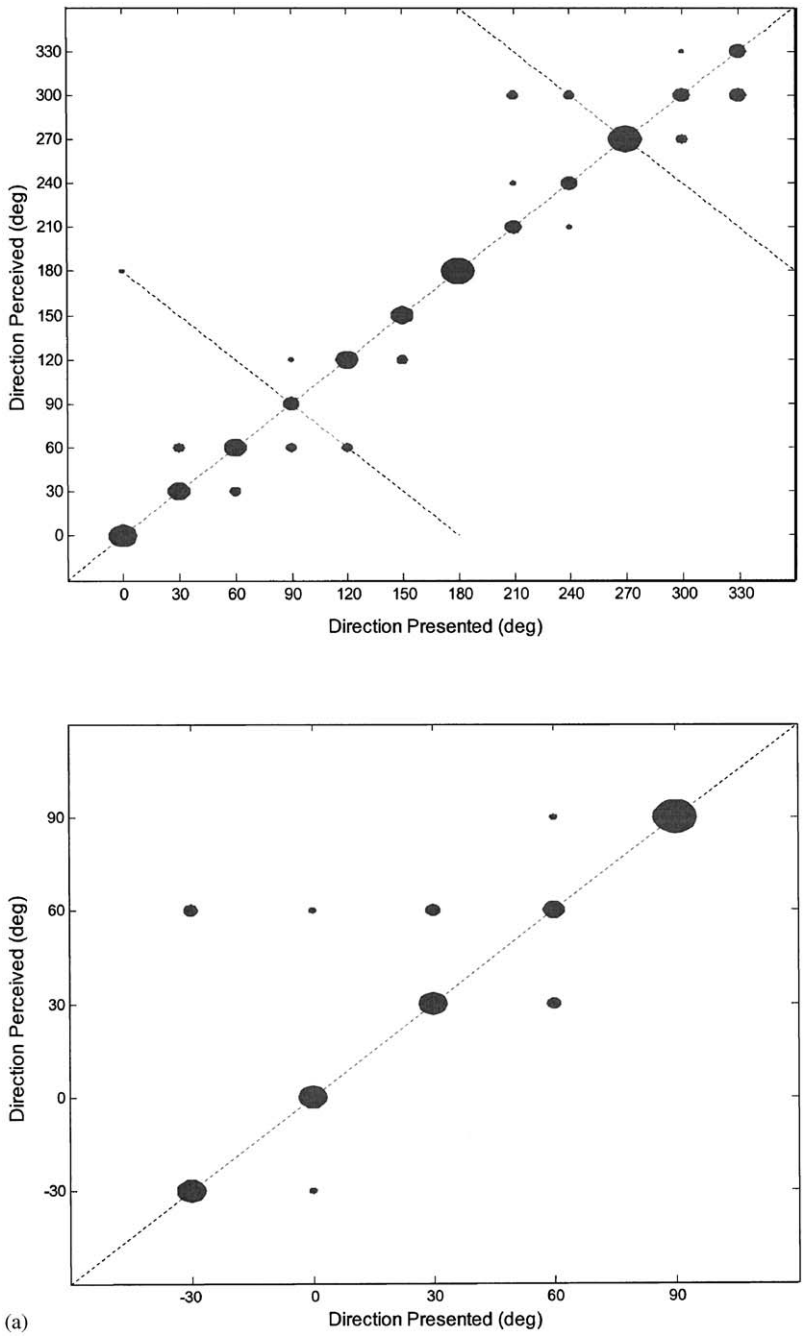


Fig. 10. (*Continued*)

where $e_{ij} = \hat{\theta}_{ij} - \theta_{ij}$ and $\hat{\theta}_{ij}$ correspond to the ith subject and the jth perceived angle, and θ_{ij} corresponds to the ith subject and they jth presented angle. From Table 1, it is concluded that the modified SVD with non-uniform sampling produced the best performance, even slightly better than direct implementation of HRTF. The performance of the simple spherical-head model was the poorest among all methods.

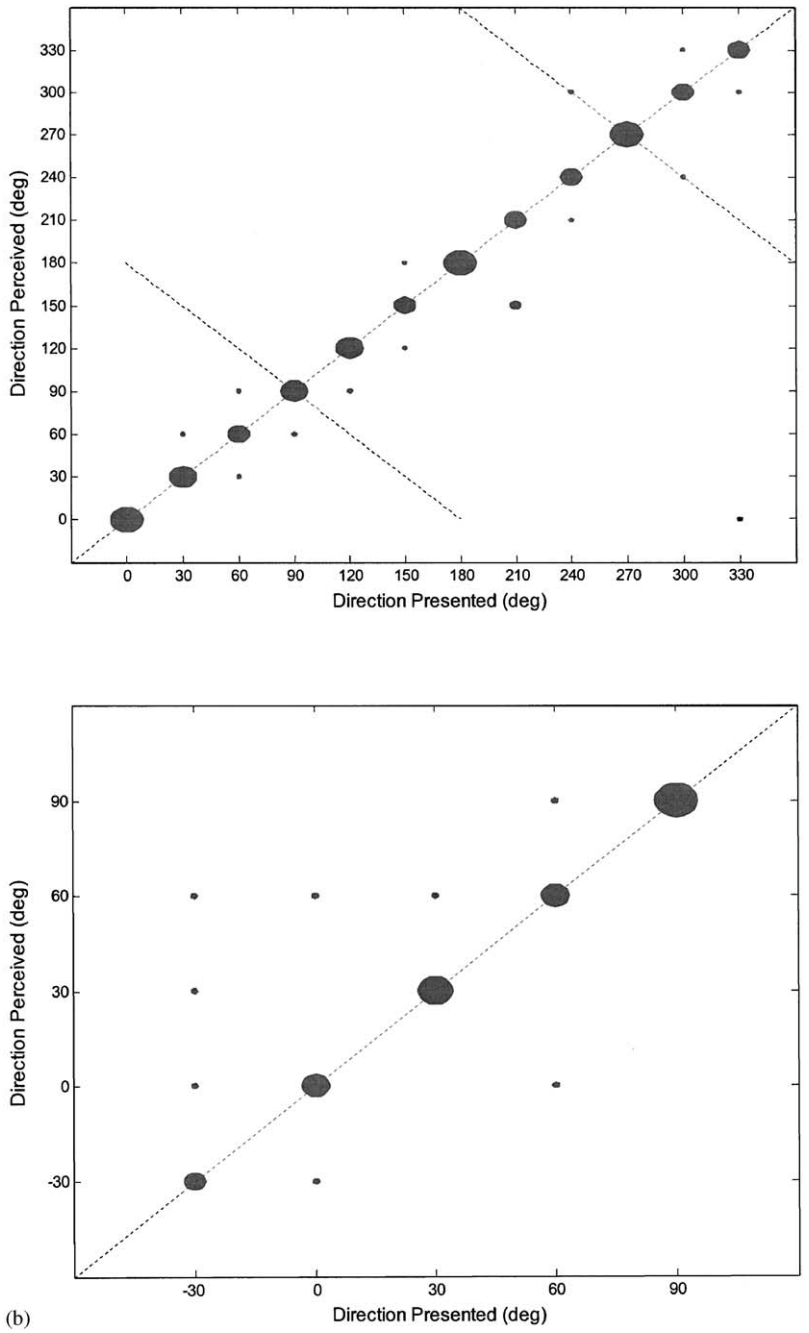
5. Conclusions

A model based on a 3D array beamformer has been developed to approximate the HRTFs which represent the spatial and temporal characteristics of the external ears. This model provides a means of synthesizing the HRTF for arbitrary broadband sources at any direction of incidence. A modified SVD procedure with regularization and non-uniform sampling technique has been employed to calculate the optimal array coefficients such that the difference between the array



(a)

Fig. 11. The result of the subjective localization experiments for the modified SVD procedure with regularization. (a) Uniform sampling approach. (b) Non-uniform sampling approach.



(b)

Fig. 11. (*Continued*)

response and the desired HRTF template in the frequency domain is minimized. The modified SVD method performed relatively well in comparison with a conventional least-square method. The non-uniform sampling scheme conforms better to the human hearing mechanism than the

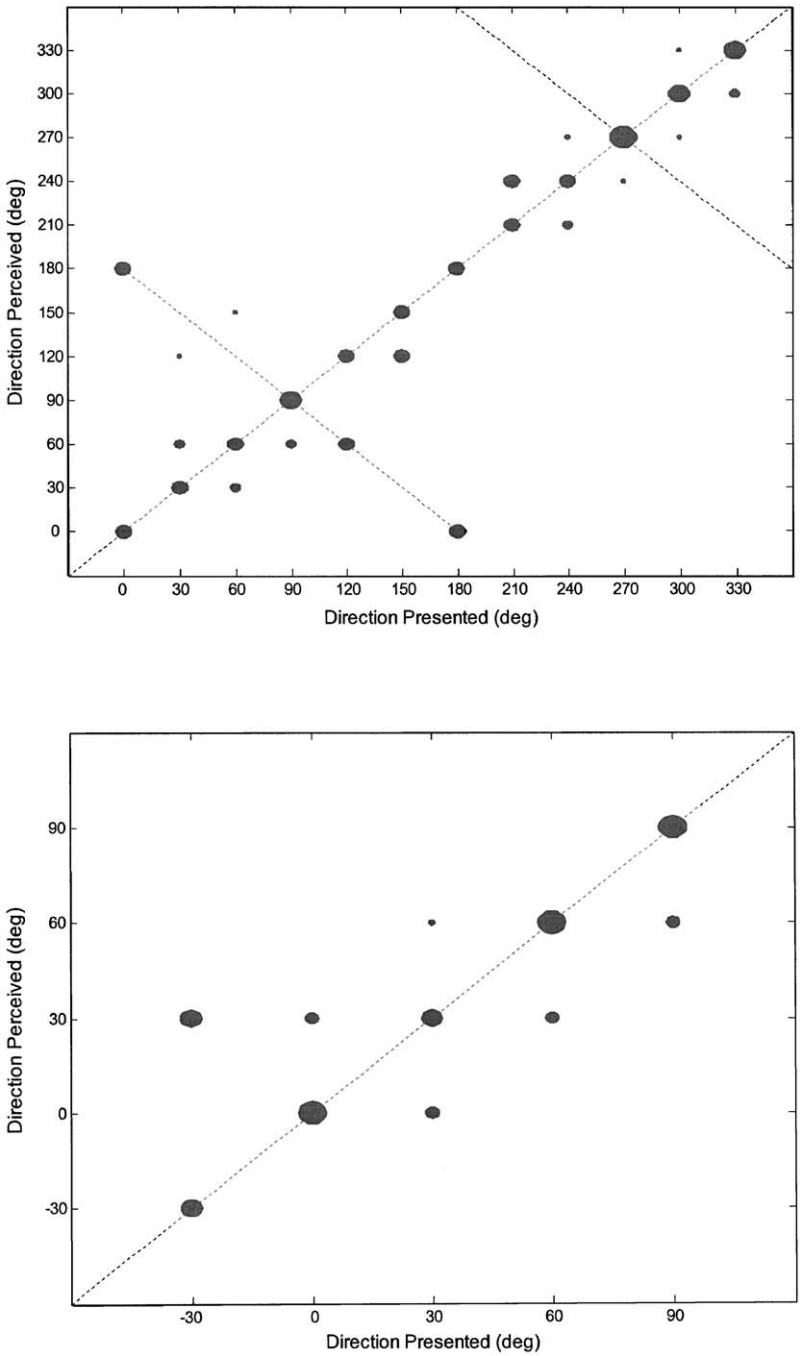


Fig. 12. The result of the subjective localization experiments for the spherical-head model.

Table 1
The mean and standard deviation of error of the array model

	μ (deg)	σ (deg)
Direct implementation of HRTF	2.67	33.54
Least-square approach	4.12	33.25
Least-square approach with non-uniform sampling	3.87	32.46
Modified SVD	3.82	31.78
Modified SVD with non-uniform sampling	2.56	21.52
Spherical head model	5.23	50.56

uniform sampling. An additional benefit of the non-uniform sampling approach is that the number of frequency samples is substantially reduced in the model-matching procedure. Apart from good modeling accuracy both in the azimuths and elevations, the proposed method has the advantages of the saving in memory storage for HRTF coefficients and the numerical stability in optimizing the array configuration using SVD. There are only several hundred array coefficients and the associated ITD needs to be stored in the synthesis of the HRTFs. From the results of the subjective localization tests, the present array beamformer model proved to be effective in approximating the original measured HRTF, without appreciable degradation of performance in localization.

Acknowledgements

The work was supported by the Nation Science Council in Taiwan, Republic of China, under the project number NSC 91-2212-E009-032.

References

[1] D.R. Begault, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, Cambridge, MA, 1994.
[2] E.M. Wenzel, M. Arruda, D.J. Kistler, F.L. Wightman, Localization using non-individualized head-related transfer functions, *Journal of the Acoustical Society of America* 94 (1993) 111–123.
[3] M. Kleiner, B.I. Dahlenbäck, P. Svensson, Auralization—an overview, *Journal of the Audio Engineering Society* 41 (1993) 861–875.
[4] B. Gardner, K. Martin, *HRTF Measurements of a KEMAR Dummy-Head Microphone*, Tech. Rep. 280, MIT Media Lab, Cambridge, MA, 1994.
[5] D. Hammershøi, Audio Eng. Soc. Conv. Preprint 4155, 1996.
[6] D.W. Batteau, The role of the Pinna in human localization, *Proceedings of the Royal Society of London Series* 168 (1968) 158–180.
[7] J. Mackenzie, J. Huopaniemi, V. Välimäki, I. Kale, Low-order modeling of head-related transfer functions using balanced model truncation, *IEEE Signal Processing Letters* 4 (2) (1997) 39–41.
[8] K. Genuit, A description of the human outer ear transfer function by elements of communication theory, *Proceedings of the 12th International Congress on Acoustics*, Toronto, Canada, 1986 (ADSTR. B6-8).
[9] R.O. Duda, CIPIC Interface Lab. 3-D audio for HCI, 2000.

[10] Y. Kahana, P.A. Nelson, Spatial acoustic mode shapes of the human pinna, *Audio Engineering Society, 109th Convention*, Los Angeles, CA, 2000, Preprint 5218.
[11] J. Chen, B.D. Van Veen, K.E. Hecox, A spatial feature extraction and regularization model for the head-related transfer function, *Journal of the Acoustical Society of America* 97 (1) (1995) 439–450.
[12] J. Chen, B.D. Van Veen, K.E. Hecox, External ear transfer function modeling: a beamforming approach, *Journal of the Acoustical Society of America* 92 (4) (1992) 1933–1944.
[13] D.J. Kistler, F.L. Wightman, A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction, *Journal of the Acoustical Society of America* 91 (1992) 1637–1647.
[14] J.C. Middlebrooks, D.M. Green, Observations on a principal components analysis of head-related transfer functions, *Journal of the Acoustical Society of America* 92 (1992) 597–599.
[15] E. Zwicker, H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin, Heidelberg, 1999.
[16] J.J. Clark, M.R. Palmer, P.D. Lawrence, A transformation method for the reconstruction of functions from nonuniformly spaced samples, *IEEE Transactions on Acoustic, Speech, and Signal Processing* 33 (4) (1985) 1151–1165.
[17] O. Kirkeby, A. Nelson, H. Hamada, F. Orduna-Bustamante, Fast deconvolution of multichannel system using regulatization, *IEEE Transactions Speech and Audio Processing* 6 (2) (1998) 189–195.
[18] B. Nobel, J.W. Daniel, *Applied Linear Algebra*, Prentice-Hall, Upper Saddle River, NJ, 2000.
[19] A.V. Oppenheim, R.W. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
[20] S.J. Elliott, *Signal Processing for Active Control*, Academic Press, London, 2001.
[21] Y. Kahana, P.A. Nelson, O. Kirkeby, H. Hamada, A multiple microphone recording technique for the generation of virtual acoustic images, *Journal of the Acoustical Society of America* 105 (1999) 1503–1516.