Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Sound and Vibration 290 (2006) 1269–1289
JOURNAL OF SOUND AND VIBRATION
www.elsevier.com/locate/jsvi

Development and implementation of cross-talk cancellation system in spatial audio reproduction based on subband filtering

Mingsian R. Bai*, Chih-Chung Lee

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

Received 2 April 2004; received in revised form 25 April 2005; accepted 16 May 2005
Available online 8 August 2005

Abstract

An integrated spatial audio system based on subband filtering and a panel speaker array is developed in this paper. This system is intended for a personal computer and is capable of rendering sound images positioned arbitrarily around a listener, synchronizing with the video image. The proposed system features the spatial audio technologies such as the head-related transfer function (HRTF), the reverberator, and the cross-talk cancellation system (CCS). Of particular importance in this paper is that inverse filtering with Tikhonov regularization technique is employed in designing the multi-channel filters to cancel the cross-talks. To ease the computation loading, the CCS is implemented for frequencies below 5.5 kHz by using a subband filtering approach. The system is implemented on a 5×1 panel speaker array. Experimental investigation indicated that the proposed system is effective in creating an immersive sound field, in complement with video rendering.

1. Introduction

A spatial audio system enables positioning sound images in arbitrary directions and distances. Immersive sensation in a three-dimensional (3D) sound field is created with the aid of computers

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

0022-460X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsv.2005.05.016

and digital signal processing. Spatial audio finds applications in virtual reality computer games, home theater, PC multimedia, flight/driving simulator, teleconferencing, and so forth. The directional cues for human hearing are embedded in the transformation of sound pressure from the free field to the ears of a listener. A head-related transfer function (HRTF) [1,2] is a measurement of such transformation for a specific sound location relative to the head, and describes the diffraction of sound by the torso, head, and external ear. A synthetic binaural signal can be created by convolving a sound with the appropriate HRTFs. Spatial audio effects are generally rendered with headphones or loudspeakers. While headphones are often used for binaural audio reproduction, they often suffer from in-head localization or front-back reversals [3], not to mention the inconvenience during wearing. An alternative way that avoids the problems of headphones is to use stereo loudspeakers for audio reproduction. Despite the benefits of loudspeaker rendering, cross-talk arises as a major problem that could adversely effect the 3D audio quality due to the Haas effect [4]. This motivates the development of cross-talk cancellation systems (CCS) that seek to minimize the influences due to cross-talks in loudspeaker reproduction.

Methods have been suggested to address the cross-talk cancellation problem. The first proposition of CCS was perhaps due to Schroeder and Atal [5], and later Damaske and Mellert [6]. Their systems were limited to a zone allowing 75–100 mm head movement beyond which the spatial sound effect would vanish. Cooper and Bauck suggested a method that modeled the head as a sphere and calculated the ipsilateral and contralateral terms [7]. A similar method by Gardner approximates the effect of the head with geometric delays and low-pass filters that account for head shadowing [8]. Cooper and Bauck [9] and Bauck and Cooper [10] suggested a simplified CCS, or the "shuffler filter," by assuming the left-right symmetry of system functions. One issue that frequently arises in practical application is the head movement of listener. To cope with the issue, head-tracking CCS have been reported [11,12]. Alternatively, as suggested by Takeuchi and Nelson, the robustness of CCS can be enhanced against head movement by closely placing two speakers to form the so-called stereo dipole [13]. An extension of this concept was the optimal source distribution (OSD) system [14].

In this work, a spatial audio system featuring HRTF, CCS, and a reverberator is integrated with the panel speaker and array signal processing technology. This system is intended for a personal computer with a single user, and is capable of rendering sound images positioned arbitrarily around the listener, synchronizing with the video image. The block diagram of the integrated system is shown in Fig. 1, wherein the HRTFs position the sound sources and down-mix into binaural signals, the reverberator simulates room effects, and the CCS cancels cross-talks. Due to space limitation, this paper focuses primarily on the development of CCS. Our CCS filters are designed, using a method parallel to the inverse filtering technique suggested by Kirkeby et al. [15]. However, the present method differs from Kirkeby's method in that a more sophisticated frequency-dependent regularization scheme is employed in the filter synthesis stage [16]. Another unique feature of the proposed system is the band-limited implementation using subband filtering. In considering the computation loading and robustness against uncertainties of HRTFs and head movement, the proposed CCS is band-limited to frequencies below 5.5 kHz [17]. To accomplish the band-limited implementation, we adopted a subband filtering technique based on a M-channel quadrature mirror filter (QMF) bank [18]. In this design, the perfect reconstruction (PR) condition is fulfilled.

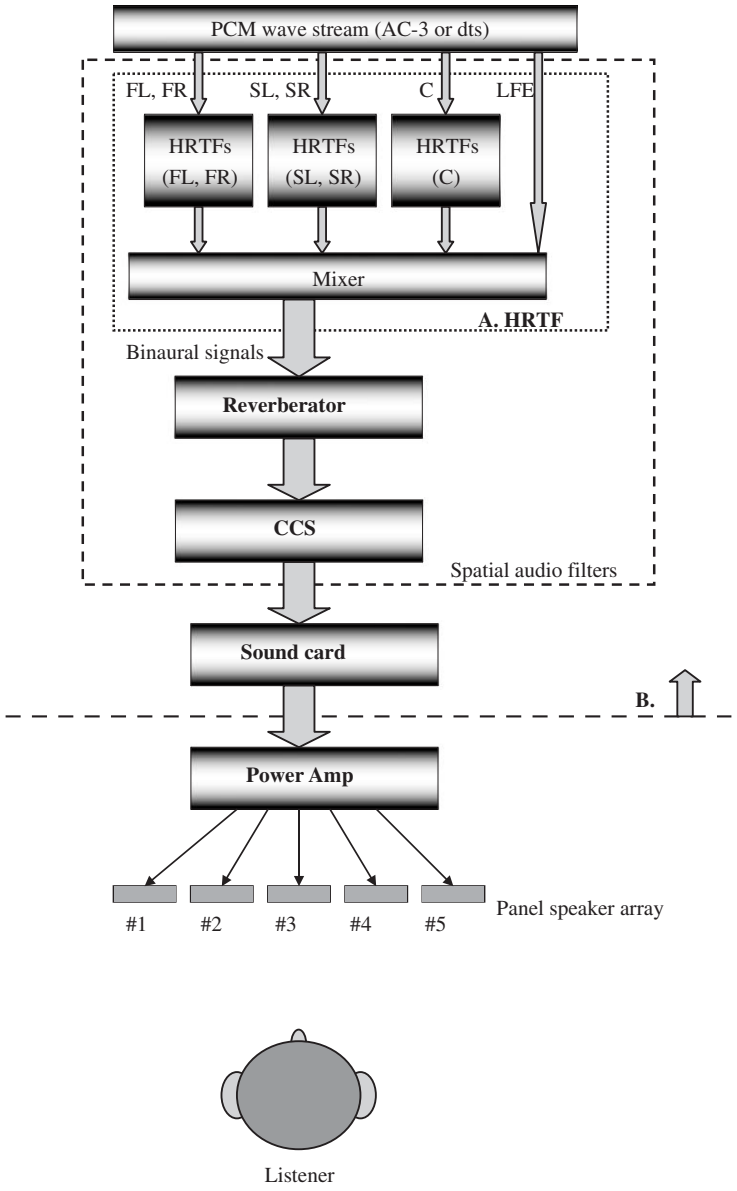

Fig. 1. The block diagram of the spatial audio system.

Instead of using conventional stereo loudspeakers, the proposed CCS is implemented on a panel speaker array. The panel speakers are light, thin and small, making them well suited for computer multimedia applications. Detailed investigation on panel speakers can be found in Ref. [19]. On the other hand, there are three reasons of using such array configuration. First, the closely spaced panel speakers provide robustness against head misalignment and compactness enabling direct placement on a computer monitor. Second, the deficiency associated with panel

speakers such as non-flat frequency response and low bass efficiency can be corrected by using an array configuration [19]. Third, a wide range of array signal processing techniques can be exploited for beam forming and steering purpose [20]. Array speakers give us more latitude in controlling the sound field in the design of a CCS. There are generally six output channels on a sound card of multimedia PC—one for subwoofer and the others can be connected to the 5×1 panel speaker array.

The proposed CCS is applied to multimedia presentation on a personal computer. Numerical simulation and experimental investigation are carried out to justify the proposed spatial audio system. Design issues and technical considerations are discussed.

2. Inverse filtering with Tikhonov regularization

As mentioned previously, a CCS aims to cancel the cross-talks in stereo loudspeaker rendering so that the binaural signals are reproduced at two ears as that from a headphone. This can be regarded as a model-matching problem shown in Fig. 2. $\mathbf{x}(z)$ is a vector of U program input signals, $\mathbf{u}(z)$ is a vector of B binaural signals, $\mathbf{v}(z)$ is a vector of S speaker input signals, $\mathbf{w}(z)$ is a vector of R reproduced signals, $\mathbf{d}(z)$ is a vector of R desired signals, and $\mathbf{e}(z)$ is a vector of R error signals. $\mathbf{M}(z)$ is an $R \times B$ matrix of matching model, $\mathbf{H}(z)$ is an $R \times S$ plant transfer matrix, and $\mathbf{C}(z)$ is an $S \times B$ matrix of CCS filters. The term z^{-m} accounts for the modeling delay to ensure causality. For the system, it is straightforward to establish the following relationships:

$$\mathbf{v}(z) = \mathbf{C}(z)\mathbf{u}(z), \tag{1}$$

$$\mathbf{w}(z) = \mathbf{H}(z)\mathbf{v}(z), \tag{2}$$

$$\mathbf{d}(z) = z^{-m}\mathbf{M}(z)\mathbf{u}(z), \tag{3}$$

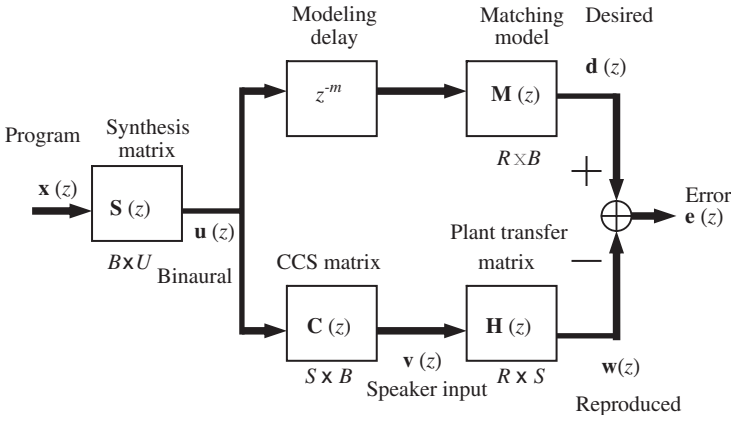

Fig. 2. The model matching problem of the CCS. $\mathbf{x}(z)$ is a vector of program input signals, $\mathbf{u}(z)$ is a vector of binaural signals, $\mathbf{v}(z)$ is a vector of speaker input signals, $\mathbf{w}(z)$ is a vector of reproduced signals, $\mathbf{d}(z)$ is a vector of desired signals, and $\mathbf{e}(z)$ is a vector of error signals. $\mathbf{M}(z)$ is a matrix the matching model, $\mathbf{H}(z)$ is the plant transfer matrix, and $\mathbf{C}(z)$ is a matrix of CCS filters.

$$\mathbf{e}(z) = \mathbf{d}(z) - \mathbf{w}(z). \tag{4}$$

Ideal model matching requires that $\mathbf{H}(z)\mathbf{C}(z) = \mathbf{M}(z)$. $\mathbf{H}(z)$ is generally non-invertible because it is usually ill-conditioned and even non-square. To overcome this difficulty, we employ the Tikhonov regularization [16] in the matrix inversion process. In the method, a frequency-domain cost function J is defined as the sum of the "performance error" $\mathbf{e}^{\mathrm{H}}\mathbf{e}$ and the "input power" $\mathbf{v}^{\mathrm{H}}\mathbf{v}$:

$$J(\mathrm{e}^{\mathrm{j}\omega}) = \mathbf{e}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{e}(\mathrm{e}^{\mathrm{j}\omega}) + \beta^2(\omega)\mathbf{v}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{v}(\mathrm{e}^{\mathrm{j}\omega}). \tag{5}$$

A regularization parameter $\beta(\omega)$ weighs the input power against the performance error. If β is too small, there will be sharp peaks in the frequency responses of the CCS filters, whereas if β is too large, the cancellation performance will be rather poor. The optimal input $\mathbf{v}_{\mathrm{opt}}(\mathrm{e}^{\mathrm{j}\omega})$ can be obtained by minimizing J

$$\mathbf{v}_{\mathrm{opt}}(\mathrm{e}^{\mathrm{j}\omega}) = [\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega}) + \beta^2(\omega)\mathbf{I}]^{-1}\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{M}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{u}(\mathrm{e}^{\mathrm{j}\omega}). \tag{6}$$

This solution always exists for $\beta \neq 0$ irrespective of the dimensions and rank of $H(\mathrm{e}^{\mathrm{j}\omega})$. Consequently, the CCS matrix can be readily identified as

$$\mathbf{C}(\mathrm{e}^{\mathrm{j}\omega}) = [\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega}) + \beta^2(\omega)\mathbf{I}]^{-1}\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{M}(\mathrm{e}^{\mathrm{j}\omega}). \tag{7}$$

In the case when the desired signals $\mathbf{d}(z)$ are identical to the binaural signals $\mathbf{u}(z)$, the matrix $\mathbf{M}(z)$ is an identity matrix of order $R = B$ and the corresponding optimal filters are given by

$$\mathbf{C}(\mathrm{e}^{\mathrm{j}\omega}) = [\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega})\mathbf{H}(\mathrm{e}^{\mathrm{j}\omega}) + \beta^2(\omega)\mathbf{I}]^{-1}\mathbf{H}^{\mathrm{H}}(\mathrm{e}^{\mathrm{j}\omega}). \tag{8}$$

While the expression in Eq. (8) may look similar to that in Ref. [15], there is a distinction in the choice of $\beta(\omega)$. In our approach, the parameter $\beta(\omega)$ is frequency dependent and constrained by a gain threshold applied to $\mathbf{C}(\mathrm{e}^{\mathrm{j}\omega})$, e.g., 12 dB. This is in contrast to the approach in Ref. [15], where a constant $\beta(\omega)$ applied to all frequencies.

Next, the frequency response matrix $\mathbf{C}(\mathrm{e}^{\mathrm{j}\omega})$ is sampled at N_c equally spaced frequencies

$$\mathbf{C}(k) = [\mathbf{H}^{\mathrm{H}}(k)\mathbf{H}(k) + \beta^2(\omega)\mathbf{I}]^{-1}\mathbf{H}^{\mathrm{H}}(k), \quad k = 1, 2, \ldots, N. \tag{9}$$

The impulse responses of the inverse filters can be obtained using inverse FFT of the frequency samples of Eq. (9) in conjunction with appropriate windowing. In order to guarantee the causality of the CCS filters, cyclic shift of the impulse response matrix is generally needed, hence the modeling delay z^{-m} in Fig. 2.

3. Band-limited implementation using the multirate approach

Band-limited implementation is chosen in this work for several reasons. First, the computation loading is too high to afford a total band (0–20 kHz) implementation. For example of the 5×1 panel speaker array considered herein, the CCS would contain 10 filters. If each filter has 1024 taps, the convolution would require 10^4 multiplications and additions per sample interval. Except for special-purpose DSP engine, real-time implementation for a total band CCS is usually prohibitive for the sampling rate commonly used in audio processing, e.g., 44.1 or 48 kHz. Second, at high frequencies, the wavelength could be much smaller than a head width. Under this circumstance, the CCS would be extremely susceptible to misalignment of the listener's head and

uncertainties involved in HRTF modeling. Third, at high frequencies, a listener's head provides natural shadowing for the contralateral paths, which is more robust than direct application of CCS. Fig. 3(a)–(c) shows the simulation results of sound pressure distribution obtained from a 6×1 speaker array at 5, 6, and 7 kHz, respectively, in the far field. The head position is indicated by a circle at the origin. The axes X and Y indicate the coordinates in cm; the bar represents sound pressure in dB. The null zone decreases with increasing frequency. For these reasons, the CCS in this study is chosen to be band-limited to 5.5 kHz (the wavelength at this frequency is approximately 6 cm). To accomplish this, a four-channel QMF bank [18] is employed to divide the total audible frequency range into subbands for CCS and direct transmission, respectively.

3.1. Theoretical background

In this section, a brief review of some techniques in multirate signal processing is given in the context of the band-limited CCS design. We begin with two fundamental operations: decimation and interpolation. Fig. 4(a) shows an M-fold decimator that produces the output sequence

$$y_D(n) = y(Mn), \tag{10}$$

where M is an integer. In frequency domain, it can be shown that the output $Y_D(e^{j\omega})$ and the input $Y(e^{j\omega})$ of the decimator are related by

$$Y_D(e^{j\omega}) = \frac{1}{M}\sum_{k=0}^{M-1} Y(e^{j(\omega-2\pi k)/M}). \tag{11}$$

On the other hand, Fig. 4(b) shows an L-fold expander that takes the input $x(n)$ and produces an output sequence

$$y_E(n) = \begin{cases} x(n/L) & \text{if } n \text{ is integer-multiple of } L, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

In frequency domain, it can be shown that the output $Y_E(e^{j\omega})$ and the input $X(e^{j\omega})$ of the expander are related by

$$Y_E(e^{j\omega}) = X(e^{j\omega L}). \tag{13}$$

In general, the decimator is preceded with a digital lowpass filter called the decimation filter and the expander is followed by a digital lowpass filter called the interpolation filter. These play similar roles as the anti-aliasing filter and reconstruction filter in analog signal processing.

It was not until the discovery of two concepts, the noble identities and polyphase representation [18], that multirate signal processing became efficient enough for practical implementation. Fig. 5(a) and (b) depict the idea of the noble identities, wherein two systems are equivalent. In the polyphase representation, on the other hand, a rational transfer function $G(z)$ can be decomposed as

$$G(z) = \sum_{n=-\infty}^{\infty} g(nM)z^{-nM} + z^{-1}\sum_{n=-\infty}^{\infty} g(nM+1)z^{-nM} + \cdots + z^{-(M-1)}\sum_{n=-\infty}^{\infty} g(nM+M-1)z^{-nM}. \tag{14}$$

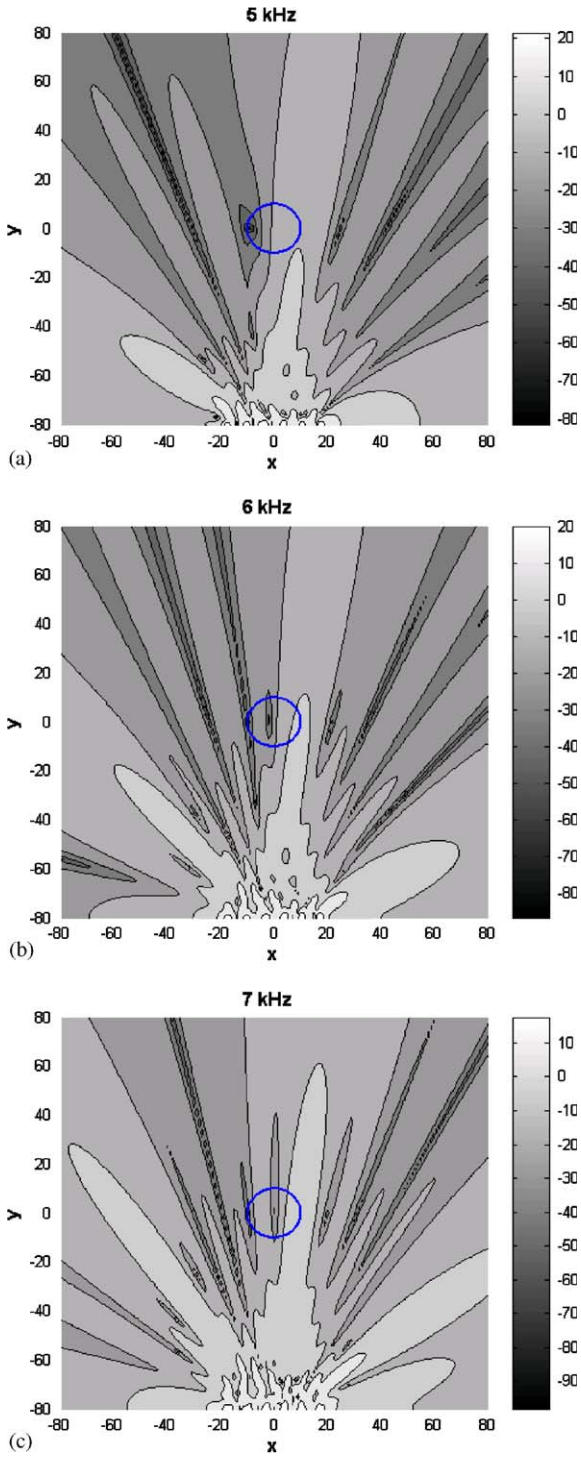

Fig. 3. The simulation results of sound pressure distribution obtained from a 6×1 speaker array in the far field. (a) 5 kHz, (b) 6 kHz, (c) 7 kHz.

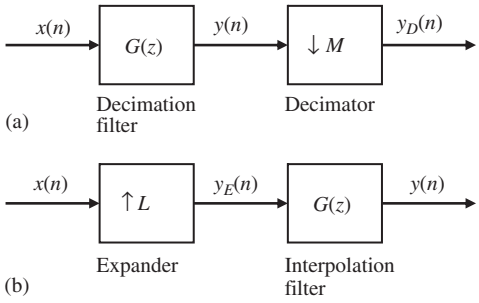

Fig. 4. Two building blocks in multirate signal processing. (a) The decimation module, (b) the interpolation module.

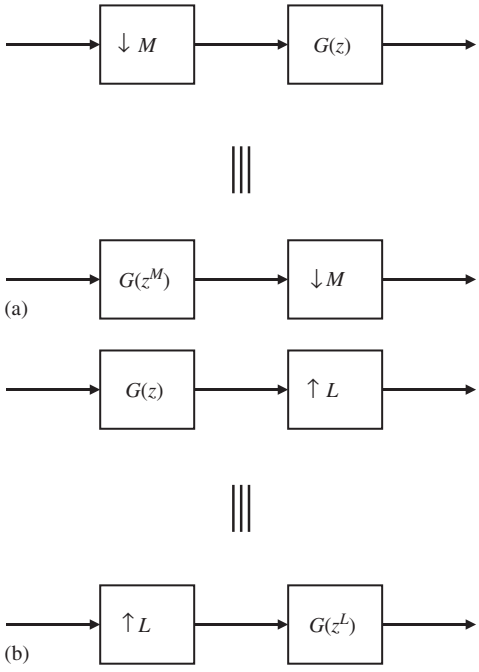

Fig. 5. The noble identities for multirate systems, (a) The first identity, (b) the second identity.

This can be compactly written as

$$G(z) = \sum_{\ell=0}^{M-1} z^{-\ell} E_\ell(z^M), \tag{15}$$

where

$$E_\ell(z) = \sum_{n=-\infty}^{\infty} e_\ell(n) z^{-n} \tag{16}$$

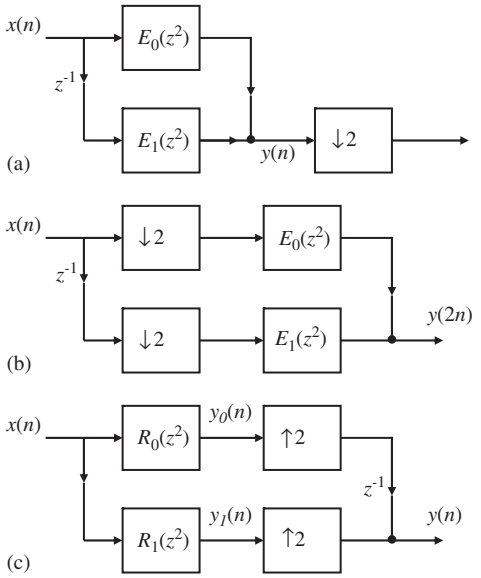

Fig. 6. The polyphase implementation. (a) Polyphase implementation of a decimation filter, (b) an equivalent implementation of the decimation filter in (a), (c) polyphase implementation of an interpolation filter.

with

$$e_\ell(n) \triangleq g(Mn+\ell), \quad 0 \leqslant \ell \leqslant M-1. \tag{17}$$

Eq. (15) is called the Type 1 polyphase representation and $E_\ell(z)$ is the polyphase component of $G(z)$. An alternative way of writing Eq. (14) is the Type 2 polyphase representation

$$G(z) = \sum_{\ell=0}^{M-1} z^{-(M-1-\ell)} R_\ell(z^M). \tag{18}$$

In fact, the Type 2 polyphase components $R_\ell(z)$ are the permutations of $E_\ell(z)$, i.e., $R_\ell(z) = E_{M-1-\ell}(z)$.

For example, consider the decimation filter shown in Fig. 4(a) with $M = 2$. Representing $G(z)$ by Eq. (15) leads to the block diagram of Fig. 6(a). By invoking the first noble identity, this can be redrawn as Fig. 6(b). It can be shown that the modified implementation is more efficient than the direct implementation of the filter $G(z)$. An alternate structure can also be obtained by using the Type 2 polyphase representation, as shown in Fig. 6(c). These representations permit great simplification of computation and efficient implementation of decimation and interpolation filters, as will be discussed next in the CCS filter bank design.

3.2. M-channel QMF bank

Fig. 7(a) shows the two-channel version of a QMF bank, wherein $G_0(z)$ and $G_1(z)$ are lowpass and highpass filters shown in Fig. 7(b). In practice, the analysis filters have non-zero transition bandwidth and stop-band magnitude. Consequently, the signals $x_k(n)$ are not perfectly

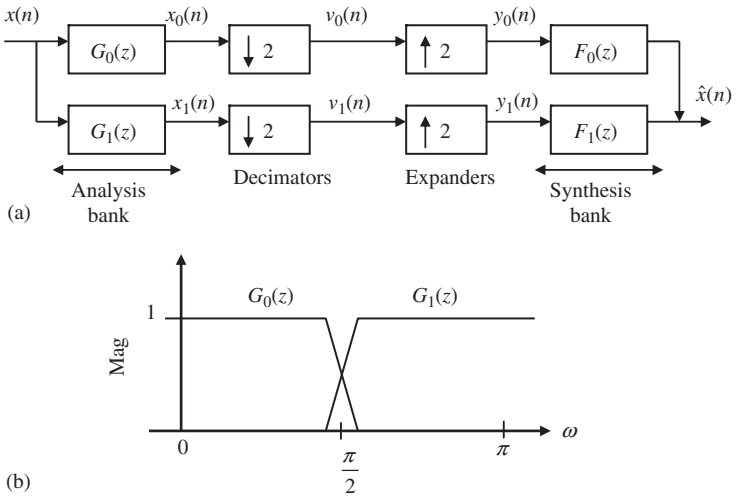

Fig. 7. The QMF bank. (a) The filter bank structure, (b) the magnitude response of the analysis filters.

band-limited, and decimation of which results in aliasing. The reconstructed signal $\hat{x}(n)$ generally differs from $x(n)$ due to three factors: aliasing, amplitude distortion, and phase distortion. It is possible that the filters can be designed in such a way that these distortions are eliminated.

From Eqs. (11) and (13), $X(z)$ and $\hat{X}(z)$ are related by

$$\hat{X}(z) = T(z)X(z) + A(z)X(-z), \tag{19}$$

where

$$T(z) = \tfrac{1}{2}[G_0(z)F_0(z) + G_1(z)F_1(z)], \tag{20}$$

$$A(z) = \tfrac{1}{2}[G_0(-z)F_0(z) + G_1(-z)F_1(z)]. \tag{21}$$

$T(z)$ and $A(z)$ are called the distortion function and aliasing function, respectively. From Eq. (21), it is clear that we can cancel aliasing by choosing the filters such that the quantity $A(z)$ is zero

$$F_0(z) = G_1(-z), \quad F_1(z) = -G_0(-z). \tag{22}$$

From Eqs. (19)–(22), Eq. (19) becomes

$$\hat{X}(z) = \tfrac{1}{2}T(z)X(z). \tag{23}$$

Assume that $G_0(z)$ is power symmetric, satisfying

$$\widetilde{G}_0(z)G_0(z) + \widetilde{G}_0(-z)G_0(-z) = 1, \tag{24}$$

where $\tilde{G}_0(z) = G_0^*(1/z^*)$. By "power symmetric," we mean that the zeros-phase filter $G(z) = \widetilde{G}_0(z)G_0(z)$ is a half-band filter, with $G(e^{j\omega})$ being non-negative. Based on Eqs. (20), (23) and (24), Eq. (23) can be reduced to $\hat{X}(z) = 0.5z^{-N}X(z)$, provided the filter $G_1(z)$ is chosen as

$$G_1(z) = -z^{-N}\widetilde{G}_0(-z) \tag{25}$$

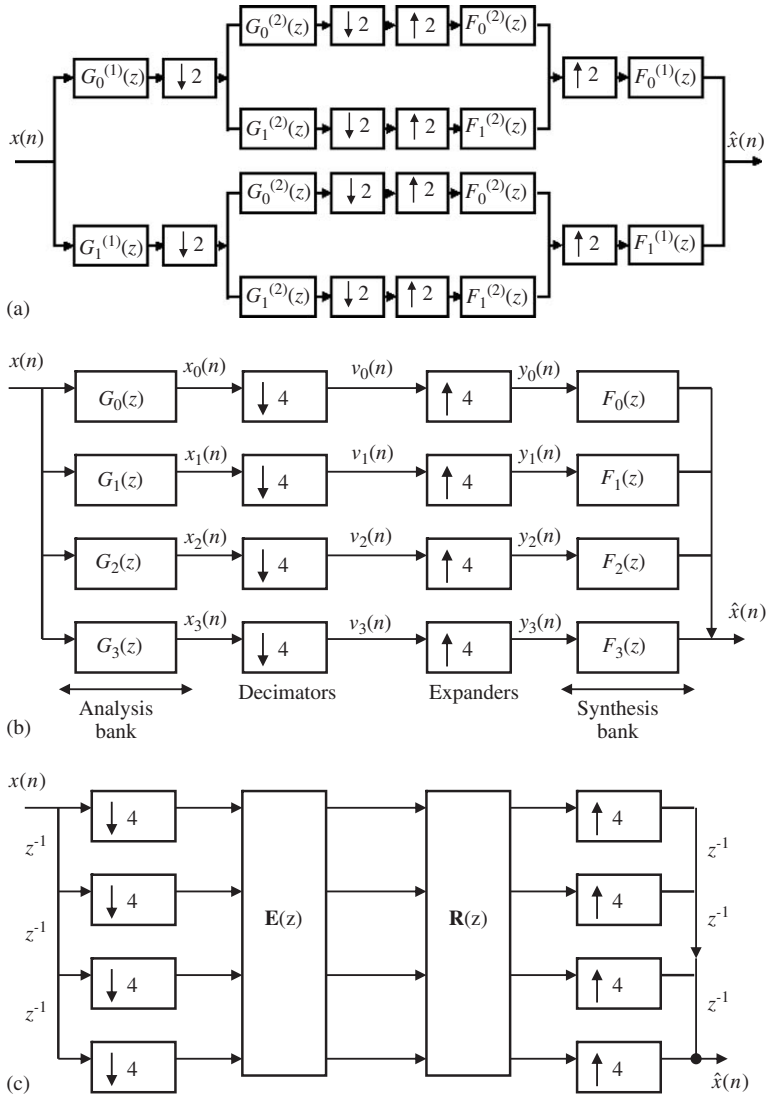

Fig. 8. The block diagram of a four-channel QMF bank. (a) A two-level maximally decimated tree structured filter bank, (b) the four-channel maximally decimated filter bank, (c) polyphase representation of a four-channel QMF bank.

for some odd integer N. Therefore, a PR system results and only the design of one filter $G_0(z)$ needs to be concerned.

Consider the structure shown in Fig. 8(a), where a signal is split into two subbands, and after decimation, each subband is again split into two and decimated. The subbands are then combined, two at a time, by using two-channel synthesis filter banks. This system is said to be a *maximally decimated tree structured filter bank* [18]. Complete system can be redrawn in an equivalent form shown in Fig. 8(b) by using the noble identities. The resulting filters $G_m(z)$ and $F_m(z)$ can be

expressed in terms of the filters $G_m^{(k)}(z)$ and $F_m^{(k)}(z)$ as follows:

$$G_0(z) = G_0^{(1)}(z)G_0^{(2)}(z^2), \quad G_1(z) = G_0^{(1)}(z)G_1^{(2)}(z^2),$$

$$G_2(z) = G_1^{(1)}(z)G_0^{(2)}(z^2), \quad G_3(z) = G_1^{(1)}(z)G_1^{(2)}(z^2),$$

$$F_0(z) = F_0^{(1)}(z)F_0^{(2)}(z^2), \quad F_1(z) = f_0^{(1)}(z)F_1^{(2)}(z^2),$$

$$F_2(z) = F_0^{(1)}(z)F_1^{(2)}(z^2), \quad F_3(z) = F_1^{(1)}(z)F_1^{(2)}(z^2). \tag{26}$$

If the two-channel system is PR, then so is the complete system.

In order to enhance computational efficiency, the Type 1 polyphase representation is used to express the transfer function $G_k(z)$ in the form

$$G_k(z) = \sum_{\ell=0}^{3} z^{-\ell} G_{k\ell}(z^4), \quad k = 0, 1, 2, 3. \tag{27}$$

In matrix form

$$\begin{bmatrix} G_0(z) \\ G_1(z) \\ G_2(z) \\ G_3(z) \end{bmatrix} = \begin{bmatrix} E_{00}(z^4) & E_{01}(z^4) & E_{02}(z^4) & E_{03}(z^4) \\ E_{10}(z^4) & E_{11}(z^4) & E_{12}(z^4) & E_{13}(z^4) \\ E_{20}(z^4) & E_{21}(z^4) & E_{23}(z^4) & E_{24}(z^4) \\ E_{30}(z^4) & E_{31}(z^4) & E_{32}(z^4) & E_{33}(z^4) \end{bmatrix} \begin{bmatrix} 1 \\ z^{-1} \\ z^{-2} \\ z^{-3} \end{bmatrix} \tag{28}$$

or

$$\mathbf{g}(z) = \mathbf{E}(z^4)\mathbf{d}_e(z). \tag{29}$$

The synthesis filters can also be expressed in a similar manner using Type 2 polyphase representation

$$F_k(z) = \sum_{\ell=0}^{3} z^{-(3-\ell)} R_{\ell k}(z^4), \quad k = 0,\ 1,\ 2,\ 3. \tag{30}$$

Using matrix notations,

$$\begin{aligned} &[F_0(z) \quad F_1(z) \quad F_2(z) \quad F_3(z)] \\ &= \begin{bmatrix} z^{-3} & z^{-2} & z^{-1} & 1 \end{bmatrix} \begin{bmatrix} R_{00}(z^4) & R_{01}(z^4) & R_{02}(z^4) & R_{03}(z^4) \\ R_{10}(z^4) & R_{11}(z^4) & R_{12}(z^4) & R_{13}(z^4) \\ R_{20}(z^4) & R_{21}(z^4) & R_{23}(z^4) & R_{24}(z^4) \\ R_{30}(z^4) & R_{31}(z^4) & R_{32}(z^4) & R_{33}(z^4) \end{bmatrix} \end{aligned} \tag{31}$$

or

$$\mathbf{f}^{\mathrm{T}}(z) = z^{-(M-1)}\tilde{\mathbf{d}}_e(z)\mathbf{R}(z^4). \tag{32}$$

An equivalent system per these two representations is shown in Fig. 8(c).

4. Experimental investigations

In order to justify the proposed integrated spatial audio system, experimental investigations were carried out. This system features a 5×1 panel speaker array, a power amplifier, a personal computer with an Intel Pentium 4 2.2 G processor, and a multi-channel sound card. The system diagram is shown in Fig. 1. The audio signal processing part is based on the HRTF, reverberator, and CCS, as mentioned previously. Our evaluation shall focus mainly on the performance of CCS. The 5×1 panel speaker array mounted on the computer monitor serves as the means for audio reproduction. The arrangement of the panel speaker array is depicted in Fig. 9(a) and (b). The size of each rectangular panel is 7 cm $\times$ 6.7 cm and the spacing between adjacent speakers is $d = 6.7$ cm. The panels are made of PU foam, with thickness 4 mm. Each panel is driven by an electromagnetic exciter affixed to an aluminum frame. The elevation of the panel speaker array is 10 cm higher than the manikin's ear. A measuring microphone is fitted inside the manikin's ear. On the other hand, in order to synchronize the audio and video data streams, Microsoft DirectShow [21] is employed for implementation of the spatial audio system. The DirectShow is exploited here as a software platform in Microsoft Windows system for computer multimedia rendering. Fig. 10 shows the photo of the complete experimental arrangement. The experiments were conducted inside an anechoic chamber, as shown in the same figure.

The CCS used in this experiment is based on the band-limited implementation. Frequency division is accomplished using a four-channel QMF bank, as detailed in Section 3. With the sampling rate 44.1 kHz, the CCS is band-limited to 5.5 kHz. Fig. 11(a) shows the frequency responses of the four-channel QMF bank. Each FIR filter has 160 taps. Fig. 11(b) shows the measured frequency response $\hat{X}(z)/X(z)$ resulting from the implementation according to Fig. 8(c).

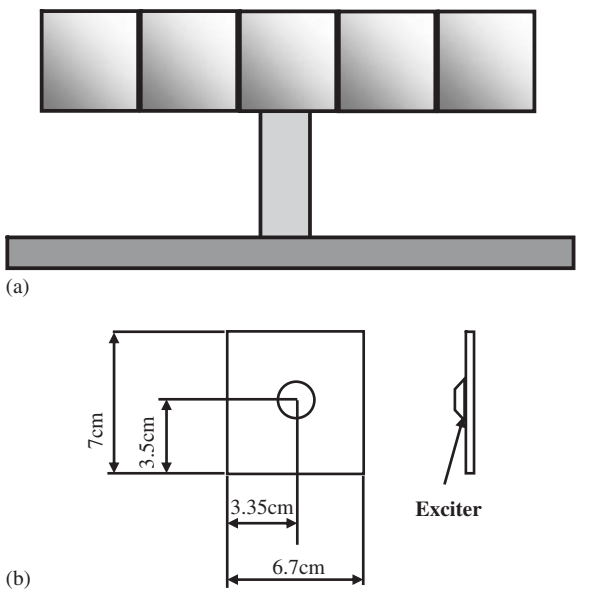

Fig. 9. The panel speaker array. (a) Arrangement of the 5×1 panel speaker array, (b) dimensions of panel speakers and the location of the exciter.

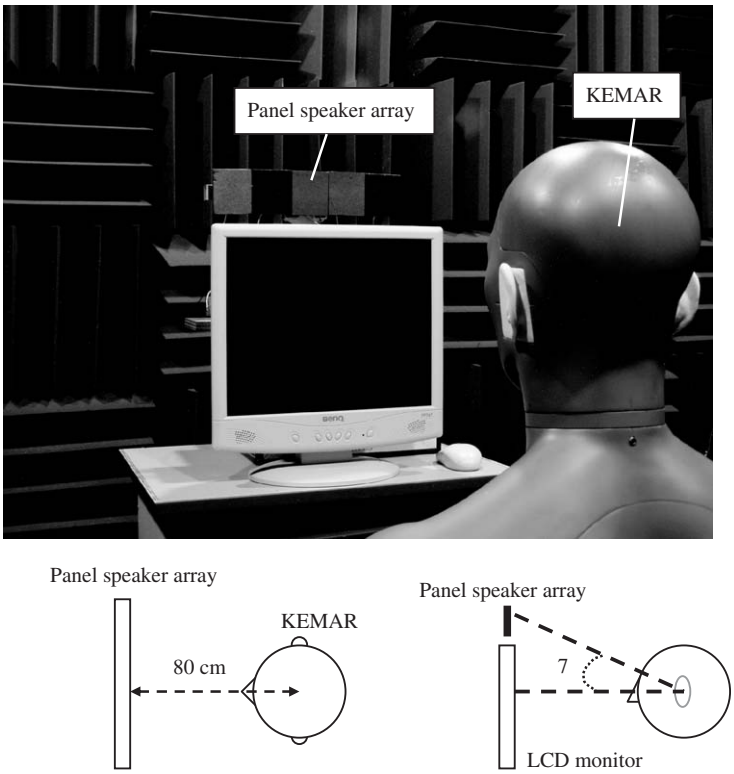

Fig. 10. The photo of the experimental arrangement of the integrated spatial audio system for a personal computer.

As evident in this result, the thus implemented filter bank is indeed a PR system since the overall system exhibit constant magnitude and linear phase.

Prior to the design of CCS, the frequency responses of the plant were measured. Since the plant has five speaker inputs and two audio outputs at ear positions, there were 10 frequency responses to be identified. Fig. 12 shows the magnitudes of the measured plant frequency responses in dB within the frequency range, 150–5.5k Hz. In the figure, the first index of the frequency response indicates the number of the panel speaker in Fig. 1, while the second index indicates the left or right ear. Notable structural resonances can be seen in the frequency responses, which is a typical feature of panel speakers. In addition, the gains at low frequencies below approximately 1 kHz tend to be somewhat low. For these plant functions, the inverse CCS filters are obtained by using Tikhonov regularization, as detailed in Section 2. Kirkeby et al. [15], applied a constant β to all frequencies. This may not adequately address the fact that the condition number of the plant matrix varies drastically with frequency. Instead, we choose in this paper a frequency-dependent β under the constraint that the magnitude responses of the inverse filters would never exceed 12 dB so as not to over-drive the loudspeakers. The value of β is calculated, under this constraint, is plotted in Fig. 13 for different frequencies. It can be observed from the result that more regularization is applied below 800 Hz than above. Beyond 800 Hz, β settles as a constant. This is not surprising since strong cross-talks exist, as reflected by ill-conditioned matrix $\mathbf{H}$, at low

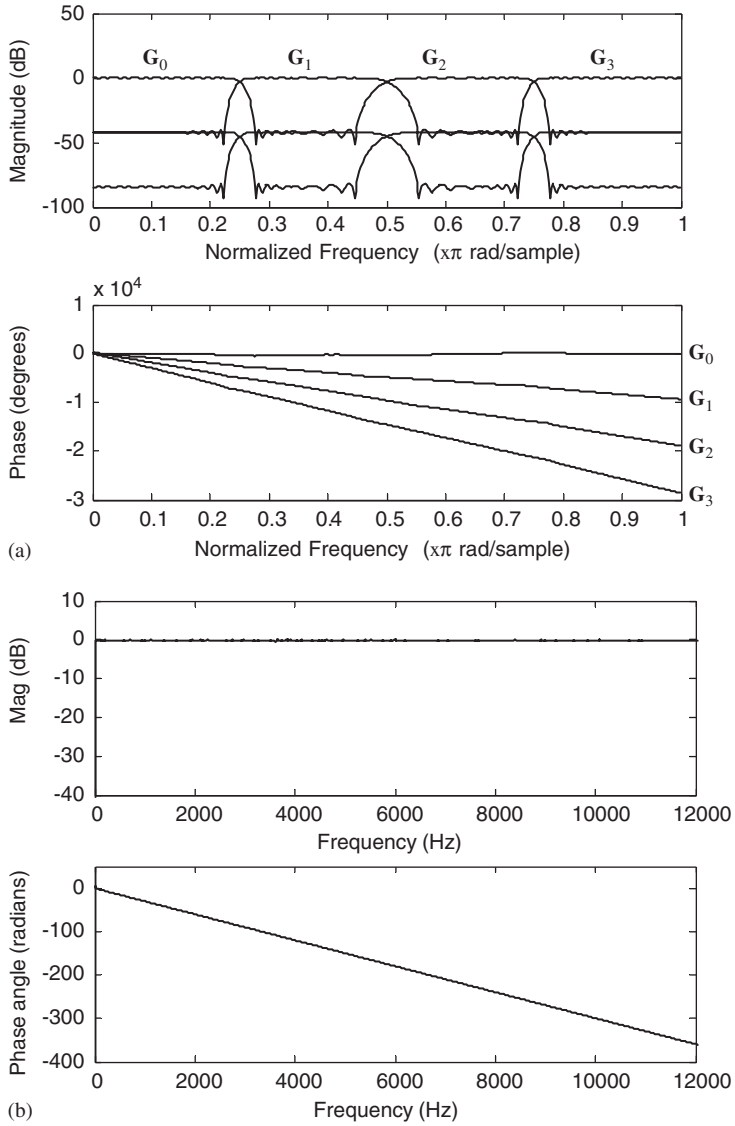

Fig. 11. The four-channel QMF bank. (a) The magnitude and phase of frequency response of each subband filter, (b) the magnitude and phase of the overall frequency response of the QMF bank, $\hat{X}(z)/X(z)$.

frequencies and overly large gains can easily result if not adequately regularized. According to this setting, the resulting frequency responses of the CCS filters are show in Fig. 14.

To facilitate the evaluation of the CCS, we further define the channel separation as the ratio of the contralateral and ipsilateral frequency responses [17]:

$$S_{ep}(\mathrm{j}\Omega) = H_{\mathrm{contra}}(\mathrm{j}\Omega)/H_{\mathrm{ipsi}}(\mathrm{j}\Omega), \tag{33}$$

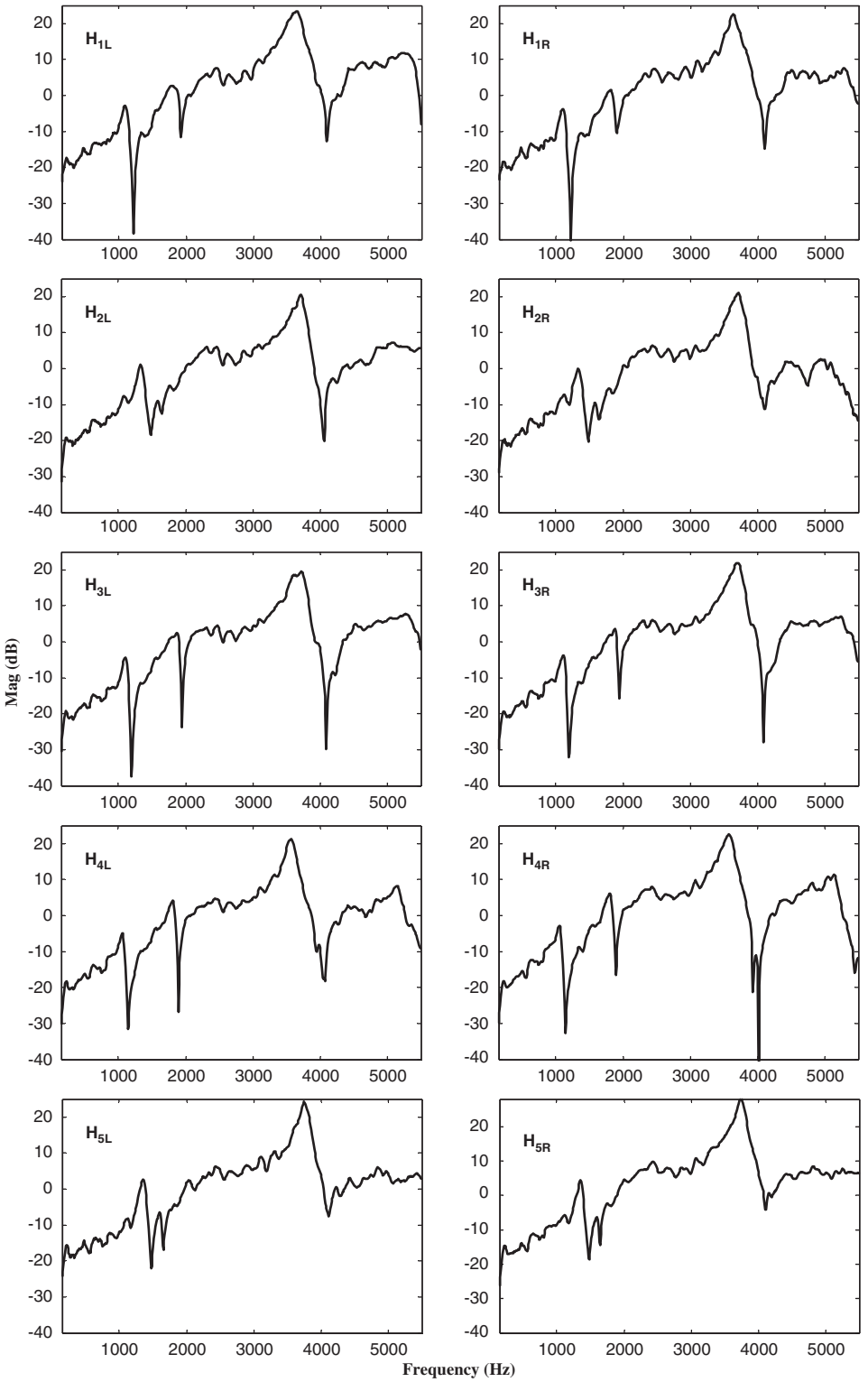

Fig. 12. The magnitudes of the plant frequency responses in dB within 150–5.5k Hz.

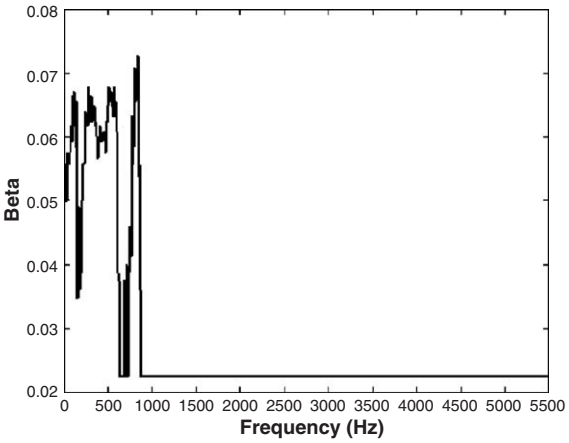

Fig. 13. The plot of β versus frequency.

where H_{ipsi} and H_{contra} symbolize the ipsilateral and contralateral frequency responses, respectively, between the loudspeakers and the ears. The smaller the value of channel separation is, the more effective the cross-talk cancellation is. Fig. 15 shows the channel separations in dB for two ears within the band 150–20k Hz. The figure on the top and bottom correspond to the measured separation at the left ear and right ear, respectively. The solid line represents the separation without CCS, or the natural channel separation. The natural channel separation generally exhibits smaller gain in high frequencies than in low frequencies due to the head shadowing effect. The dotted lines represent the results obtained using the proposed band-limited CCS. It is clear from this experimental result that the performance in terms of channel separation resulting from the CCS is rather significant in the band 1k–5.5k Hz. The maximum channel separation attains approximately 30 dB. The poor cancellation performance below 1 kHz may be attributed to the strong diffraction effect and poor loudspeaker response at that frequency range. In addition, the overall performance of the CCS can be better illustrated by examining the matrix product, $\mathbf{P} = \mathbf{HC}$, as shown in Fig. 16. Figures on the top and bottom correspond to the magnitude responses of the matrix $\mathbf{P}$ at the left ear and the right ear, respectively. The solid and the dotted lines represent the numerical and experimental results, respectively. As we can see from these plots, the matrix $\mathbf{P}$ is diagonal-dominant with relatively flat magnitude response throughout the band 1.5k–5.5k Hz. Within this control bandwidth of CCS, the system attempts to approach the identity matching model used in the CCS design. As an additional benefit of CCS, the imperfection of the panel speaker response has been compensated to render even better sound quality than the uncompensated system. In practice, however, it is impossible to achieve perfect cancellation because the plant is non-invertible and the CCS is based on approximated inverse filters. The trend of the experimental results is in good agreement with the numerical simulation, except at high frequencies. The discrepancy between the numerical and experimental results could be due to the poor signal-to-noise ratio at those frequencies.

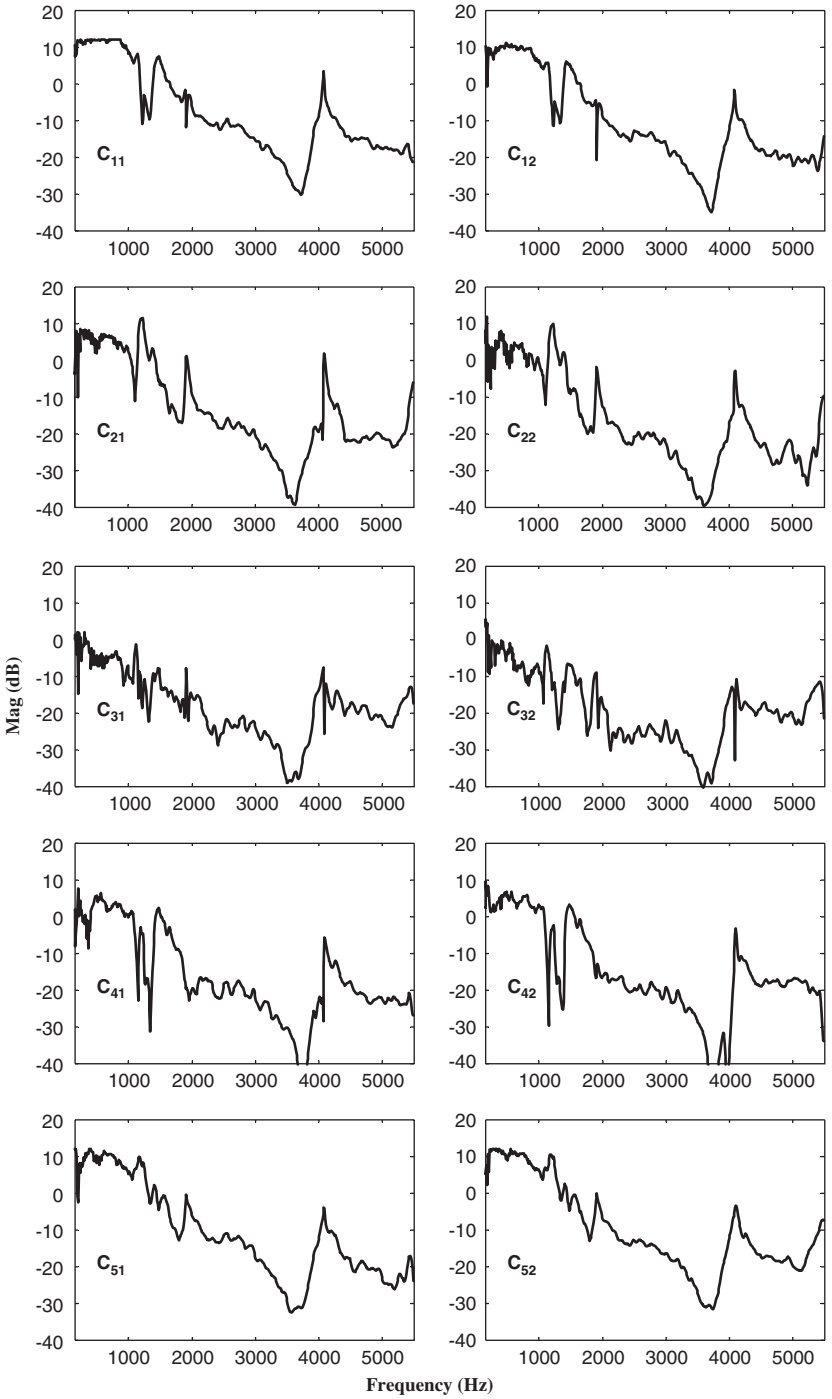

Fig. 14. The frequency responses of the CCS filters. The x-axis is the frequency in Hz and the y-axis is the magnitude in dB.

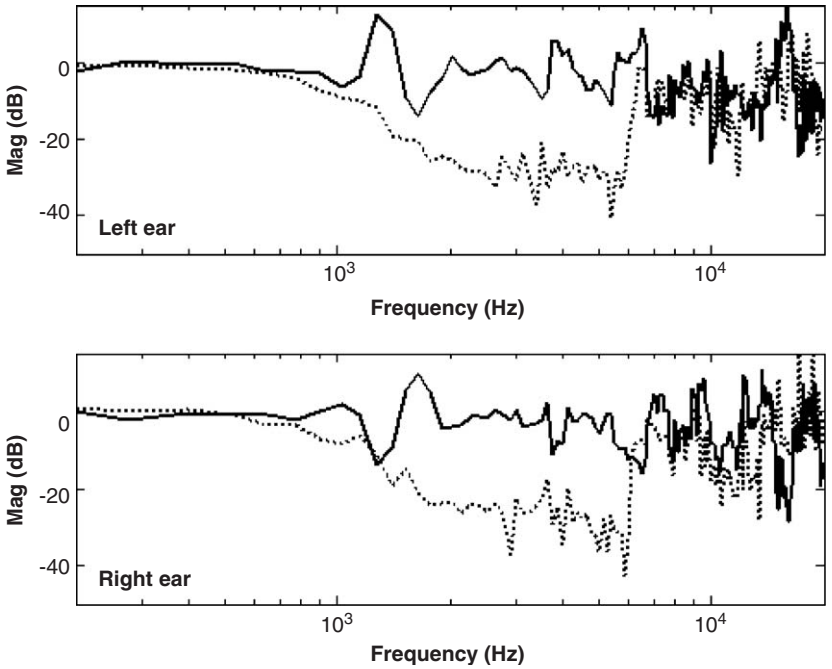

Fig. 15. The channel separations in dB for two ears within the band 150–20k Hz. The figure on the top and bottom correspond to the measured separation at the left ear and right ear, respectively (solid line: natural channel separation; dotted line: channel separation obtained using band-limited CCS).

5. Conclusions

A spatial audio system based on panel speaker array has been implemented on a personal computer. It is capable of rendering sound images positioned arbitrarily around a listener in synchronization with the video image, providing a useful solution for PC multi-media. Unlike previous systems using stereo loudspeakers, a panel speaker array is employed in this system for its compactness and robustness. The HRTF, reverberator, and CCS are all integrated in one unit. In particular, the last item is accomplished by using inverse filtering in conjunction with Tikhonov regularization. A dynamic scheme in adjusting the regularization parameter β has been proposed in this paper under a speaker input constraint. Such approach suits better the frequency-dependent needs for regularization. As indicated in the experimental results, the band-limited implementation of CCS proved to be effective in canceling the cross-talks within the control bandwidth, with reduced amount of computation loading.

Numerous limitations of the present system are pointed out as follows. First, the computation loading remains an issue for audio/video rendering. Although real-time implementation of this system is possible by using a P4 2.2 G CPU, it takes up almost 50% of the computational power when all effects (HRTF + reverberator + CCS) are enabled. Second, head misalignment of the listener remains the primary factor that affects the performance as well as robustness of the CCS, particularly at high frequencies. Methods, either fixed type or adaptive type, are currently being

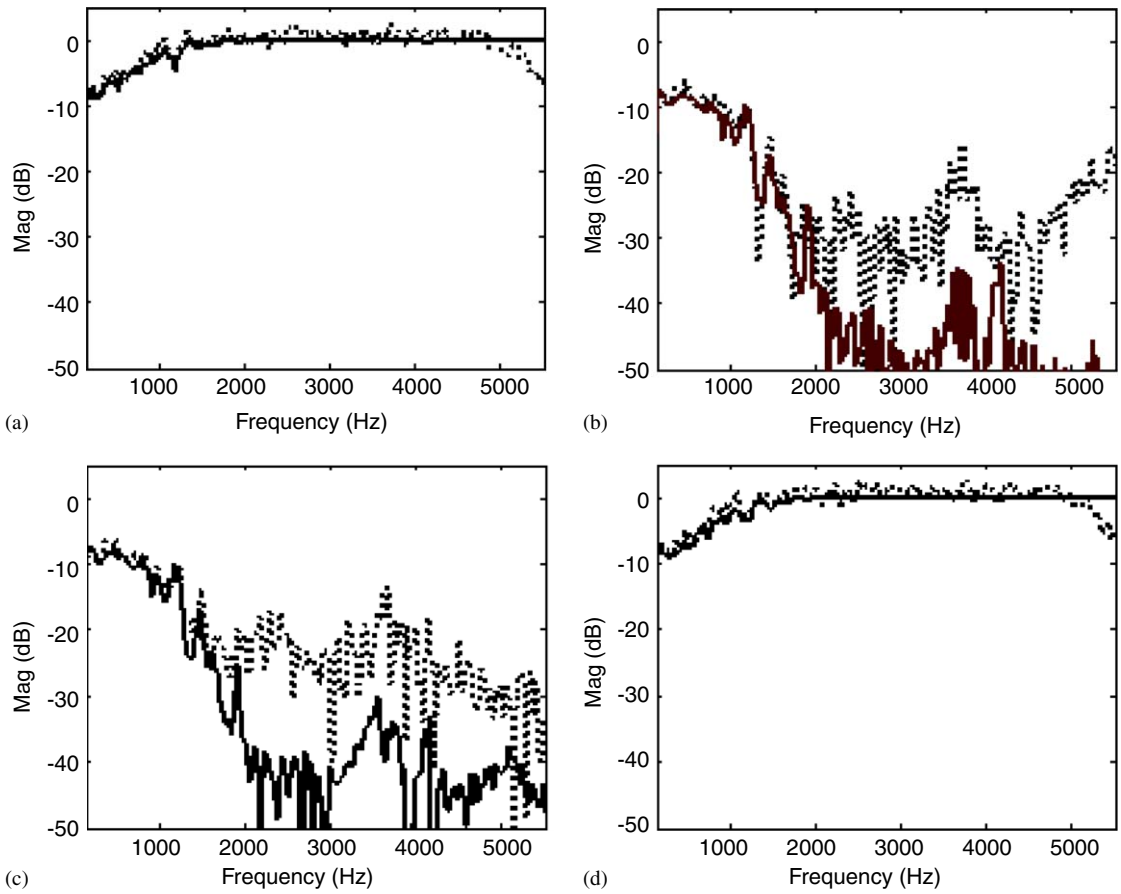

Fig. 16. The magnitudes of frequency responses of the matrix $\mathbf{P} = \mathbf{HC}$ (solid line: numerical result; dotted line: experimental result). (a) Between the left ear and the left binaural signal (P_{11}), (b) between the left ear and the right binaural signal (P_{12}), (c) between the right ear and the left binaural signal (P_{21}), (d) between the right ear and the right binaural signal (P_{22}).

sought to address this problem. Future research will focus on these aspects to enhance the practicality of the spatial audio system.

Acknowledgements

The work was supported by the National Science Council in Taiwan, ROC, under the project number NSC91-2212-E009-032.

References

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

[2] V.R. Algazi, R.O. Duda, D.M. Thompson, The CIPIC HRTF database, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.
[3] D.R. Begault, Challenges to the successful implementation of 3-D sound, *Journal of the Audio Engineering Society* 39 (1990) 864–870.
[4] A. Sibbald, Transaural acoustic crosstalk cancellation, Sensaura White Papers, 1999 (http://www.sensaura.co.uk).
[5] R. Schroeder, B.S. Atal, Computer simulation of sound transmission in rooms, *IEEE International Convention Record* 7 (1963) 150–155.
[6] P. Damaske, V. Mellert, A procedure for generating directionally accurate sound images in the upper-half space using two loudspeakers, *Acoustica* 22 (1969) 154–162.
[7] D.H. Cooper, Calculator program for head-related transfer functions, *Journal of the Audio Engineering Society* 30 (1982) 34–38.
[8] W.G. Gardner, Transaural 3D audio, MIT Media Laboratory Technical Report 342, 1995.
[9] D.H. Cooper, J.L. Bauck, Prospects for transaural recording, *Journal of the Audio Engineering Society* 37 (1989) 3–19.
[10] J.L. Bauck, D.H. Cooper, Generalized transaural stereo and applications, *Journal of the Audio Engineering Society* 44 (1996) 683–705.
[11] C. Kyriakakis, T. Holman, J.S. Lim, H. Homg, H. Neven, Signal processing, acoustics, and psychoacoustics for high-quality desktop audio, *Journal of Visual Communication and Image Representation* 9 (1997) 51–61.
[12] C. Kyriakakis, Fundamental and technological limitations of immersive audio systems, *IEEE Processing* 86 (1998) 941–951.
[13] T. Takeuchi, P.A. Nelson, Robustness to head misalignment of virtual sound imaging systems, *Journal of the Audio Engineering Society* 109 (2001) 958–971.
[14] T. Takeuchi, P.A. Nelson, Optimal source distribution for binaural synthesis over loudspeakers, *Journal of the Audio Engineering Society* 112 (2002) 2786–2797.
[15] O. Kirkeby, P.A. Nelson, H. Hamada, Fast deconvolution of multichannel systems using regularization, *IEEE Transactions on Speech and Audio Processing* 6 (1998) 189–194.
[16] A. Schuhmacher, J. Hald, Sound source reconstruction using inverse boundary element calculations, *Journal of the Acoustical Society of America* 113 (2003) 114–127.
[17] W.G. Gardner, *3-D Audio Using Loudspeakers*, Kluwer Academic, London, 1998.
[18] P.P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[19] M.R. Bai, T. Huang, Development of panel loudspeaker system: design evaluation and enhancement, *Journal of the Acoustical Society of America* 109 (2001) 2751–2761.
[20] D.H. Johnson, D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[21] Microsoft DirectShow Documentation in MSDN Library (http://msdn.microsoft.com).