Comparative study of audio spatializers for dual-loudspeaker mobile phones

Mingsian R. Bai,[a] Geng-Yu Shih, and Chih-Chung Lee
Department of Mechanical Engineering, National Chiao-Tung University, 1001 Ta-Hsueh Road, Hsin-Chu 300, Taiwan, Republic of China

(Received 16 May 2006; revised 11 October 2006; accepted 11 October 2006)

MPEG-1, layer 3 handsets equipped with dual loudspeakers and three-dimensional audio modules have received much attention in the market of consumer electronics. To create spatial impression during audio reproduction, the head-related transfer function (HRTF) and the crosstalk cancellation system (CCS) are key elements in many audio spatializers. However, there are many factors that one should take into account during the design and implementation stages of an audio spatializer in the handset application. In the paper, a comprehensive study was undertaken to compare various audio spatializers for use with dual-loudspeaker handsets, in the context of inverse filtering strategies. Two deconvolution approaches, the frequency-domain method and the time-domain method, are employed to design the required inverse filters. Different approaches to design audio spatializers with the HRTF, CCS, and their combination are compared. In particular, two modified CCS approaches are suggested. Issues in the implementation phase such as regularization, complex smoothing, and structures of inverse filters are also addressed in the paper. Comprehensive objective and subjective tests were conducted to investigate the aforementioned aspects of audio spatializers. The data obtained from the subjective tests are processed by using the multianalysis of variance to justify statistical significance of the results. © *2007 Acoustical Society of America.*
[DOI: 10.1121/1.2387121]

PACS number(s): 43.60.Dh, 43.60.Pt, 43.60.Qv, 43.60.Uv [EJS] Pages: 298–309

I. INTRODUCTION

Thanks to rapid advances of mobile communication technology, handsets have swiftly entered everyone's daily life. In addition to a simple phone, a nowadays' handset has to serve also as a camera, a personal digital assistant, MPEG-1, layer 3 (MP3) player, and even a video player in the third-generation application. In order to cater to the ever-increasing demands of high quality audio, three-dimensional (3D) audio reproduction for use with dual-loudspeaker handsets has emerged. In 3D audio reproduction, the head-related transfer function (HRTF) and the crosstalk cancellation system (CCS) are two core technologies. HRTF is a mathematical model representing the propagation process from a sound source to the human ears. HRTFs thus contain localization cues as a result of the propagation delay and the diffraction effects due to the head, ears, and even torso. This allows us to create a directional impression by properly synthesizing HRTFs at the prescribed direction.[1] Although this is effective in headphone reproduction, a crosstalk problem arises when loudspeakers are used as the rendering transducers.[2,3] To overcome this problem, the CCS based on inverse filtering are employed to minimize the effects due to crosstalk that can obscure sound image. In general, two types of deconvolution approaches, the frequency-domain method[4] and the time-domain method,[5,6] can be utilized to design the required inverse filters. Since the acoustic systems, or plants, are usually noninvertible, some regularization measures have to be taken in these methods to avoid excessive boosts for the inverse filters caused by overcompensating the acoustic system. As an effective alternative, excessive gain of the inverse filters can also be avoided by smoothing the frequency response functions of the acoustic system prior to the inversion process.[7]

In inverse filter design, Norcross *et al.* pointed out that the time-domain methods are subjectively more robust but computationally less efficient than the frequency-domain method.[8] The main difficulty in the inversion process lies in the fact that the acoustic plants are typically nonminimum phase, meaning that a causal inverse filter does not exist.[9] To cope with the problem, a modeling delay was first introduced by Clarkson *et al.*[10] Furthermore, Kirkeby *et al.*[11] used the least-squares method along with a modeling delay to find the causal inverse filters. Wang and Pai also applied the time-domain method to determine the optimal modeling delay for the inverse filters.[12]

Conventional inverse filtering leads to reduced crosstalk and equalized ipsilateral response. However, if the CCS is inadequately designed, the latter effect can result in audible high-frequency artifacts. To address the problem, two modified CCS are proposed in this paper. The idea underlying these modified methods is to eliminate the crosstalk of the contralateral paths from the loudspeakers to the listener's ears without equalizing the ipsilateral paths. The modified CCS methods also have a desirable property that the CCS is loudspeaker independent. Extensive tests were conducted in the work to compare different approaches of audio spatializers based on the HRTF, CCS, and their combination.

[a]Author to whom correspondence should be addressed; electronic mail: msbai@mail.nctu.edu.tw

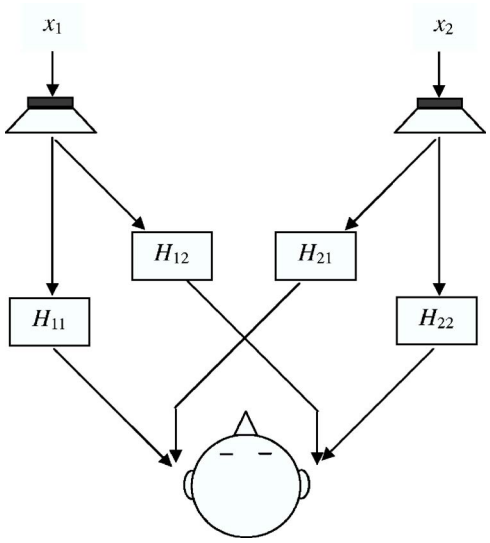

FIG. 1. Schematic diagram showing an audio reproduction system using two-channel stereo loudspeakers. Acoustic transfer functions between the loudspeakers and the listener's ears are indicated in the figure.

Another issue concerning the implementation phase is the structures of inverse filters. In considering psychoacoustic aspects and computational cost, the CCS can be implemented in a few different ways. Three structures of CCS are compared in this paper: the direct filtering method, the filter bank method,[13] and the simple lowpass mixing method.[14] The direct filtering method can be further divided into the full-band and the band-limited design.[3] The band-limited design limits the crosstalk cancellation to function only within the band 200–6 kHz.

In this work, comprehensive objective and subjective tests were conducted to investigate the aforementioned aspects of audio spatializers for mobile phones. The data of subjective tests are processed by using the multianalysis of variance (MANOVA) to justify the statistical significance of the results.[15]

II. CROSSTALK CANCELLATION SYSTEMS

A. Problem of crosstalk cancellation

Figure 1 shows a two-channel loudspeaker reproduction scenario, where H_{11} and H_{22} are ipsilateral transfer functions, and H_{12} and H_{21} are contralateral transfer functions from the loudspeakers to the listener's ears. The contralateral transfer functions, also known as the crosstalk, interfere with human's localization of sound sources when the binaural signals are reproduced by loudspeakers. In order to mitigate the effects of crosstalk, the crosstalk canceller is chosen to be the inverse of the acoustic plants such that the overall response becomes a diagonalized and distortionless response

$$\begin{bmatrix} \delta(n-m) & 0 \\ 0 & \delta(n-m) \end{bmatrix} = \begin{bmatrix} h_{11}(n) & h_{12}(n) \\ h_{21}(n) & h_{22}(n) \end{bmatrix} \otimes \begin{bmatrix} c_{11}(n) & c_{12}(n) \\ c_{21}(n) & c_{22}(n) \end{bmatrix}, \tag{1}$$

where $\otimes$ denotes convolution operation and $h_{ij}(n)$, $c_{ij}(n)$, and $\delta(n-m)$ represent the impulse responses of the respective acoustic paths, the inverse filters, and the discrete delta function delayed by m samples of delay to ensure a causal inverse filter. On the basis of inverse filtering, two deconvolution schemes along with regularization techniques are described in the following.

B. Multichannel inverse filtering with regularization

1. Frequency-domain deconvolution

The first method to be considered is the frequency-domain method[4] suggested by Kirkeby *et al.* In this method, a cost function J is defined as the sum of the "performance error" $\mathbf{e}^H\mathbf{e}$ and the "input power" $\mathbf{v}^H\mathbf{v}$,

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

with ω being the angular frequency. A regularization parameter $\beta(\omega)$ which varies from zero to infinite weighs the input power against the performance error. This is a well known Tikhonov regularization procedure. The optimal inverse filters obtained by minimizing J can be written in terms of discrete frequency index k as follows:

$$\mathbf{C}(k) = [\mathbf{H}^H(k)\mathbf{H}(k) + \beta(k)\mathbf{I}]^{-1}\mathbf{H}^H(k), \quad k = 1,2,\ldots,N_c, \tag{3}$$

where N_c-point fast Fourier transform (FFT) is assumed, and $\mathbf{H}(k)$ is the transfer matrix of acoustic plant. The coefficients of inverse filters can be obtained using the inverse FFT of the frequency response in Eq. (3), with the aid of appropriate windowing. In order to ensure the causality of the CCS filters, circular shift ($N_c/2$ maximum) of the resulting impulse response is needed to introduce a modeling delay.[16]

2. Time-domain deconvolution

The time-domain method is based on a matrix formalism of Eq. (1). In this method, a single-channel inverse filter can be obtained by solving the following matrix equation:[5,6]

$$\begin{bmatrix} d(0) \\ \vdots \\ d(N_h+N_c-2) \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} h(0) & & 0 \\ \vdots & \ddots & \vdots \\ h(N_h-1) & \ddots & h(0) \\ \vdots & \ddots & \vdots \\ 0 & & h(N_h-1) \\ \varepsilon & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varepsilon \end{bmatrix} \times \begin{bmatrix} c(0) \\ \vdots \\ c(N_c-1) \end{bmatrix}, \tag{4}$$

or simply

$$\mathbf{d} = \mathbf{hc}. \tag{5}$$

In the preceding two equations, the vector $\mathbf{d}$ represents the desired response, the matrix $\mathbf{h}$ is composed of the impulse responses $h(n)$ of acoustical plants measured *a priori*, N_h is the length of the plant impulse response $h(n)$, the vector $\mathbf{c}$ represents the impulse response of the inverse filters, and N_c

is the length of the inverse filter. The parameter ε in the lower part of the matrix $\mathbf{h}$ is a small regularization constant. The forgoing single-channel deconvolution technique can be readily extended to the two-channel case described by the following matching matrix:

$$\begin{bmatrix} \mathbf{d} \\ 0 \\ 0 \\ \mathbf{d} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_{11} & 0 & \mathbf{h}_{12} & 0 \\ 0 & \mathbf{h}_{11} & 0 & \mathbf{h}_{12} \\ \mathbf{h}_{21} & 0 & \mathbf{h}_{22} & 0 \\ 0 & \mathbf{h}_{21} & 0 & \mathbf{h}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{c}_{11} \\ \mathbf{c}_{12} \\ \mathbf{c}_{21} \\ \mathbf{c}_{22} \end{bmatrix}, \tag{6}$$

where $\mathbf{h}_{ij}$ and $\mathbf{c}_{ij}$ represent the matrices composed of the impulse responses $h_{ij}(n)$ and the coefficient vectors of the filters $c_{ij}(n)$.

The size of the matrix in Eq. (6) can be quite large. Instead of brute-force inversion, more efficient iteration techniques are employed in the work. By exploiting these properties one may use the iterative algorithms such as steepest descent and conjugate-gradient (CG) method to calculate the solution.[6] In both methods, a residual vector $\mathbf{R}$ is defined as

$$\mathbf{R} = \mathbf{D}_t - \mathbf{H}_t \mathbf{C}_t, \tag{7}$$

where $\mathbf{D}_t$, $\mathbf{H}_t$, and $\mathbf{C}_t$, represent the matrices in Eq. (6). In the steepest descent algorithm, the recursive relation for updating the coefficient of the inverse filters can be described as

$$\mathbf{C}_t(i+1) = \mathbf{C}_t(i) + \mu \mathbf{g}(i), \tag{8}$$

where i is the iterative index and $\mathbf{g}$ is the gradient vector of the cost function with a step size μ. Unlike the steepest descent algorithm, a plane search strategy based on the linear combination of gradient vectors consecutive iterations is used in the CG algorithm. Specifically, the coefficient update equation is given as

$$\mathbf{C}_t(i+1) = \mathbf{C}_t(i) + \mu \mathbf{g}(i) + \alpha \mathbf{s}(i), \tag{9}$$

where $\mathbf{s}$ is the gradient vector in last iteration and α is another step size parameter. In general, the convergence behavior of the CG method is superior to the steepest descent method due to the plane search nature of the former approach.

3. Generalized complex smoothing techniques

Due to the ill-conditioned nature of the acoustical system, how to properly limit the gain of the inverse filter is a critical issue in designing the CCS. One way to deal with this problem is the regularization method, as already mentioned in the previous section. Another simple but elegant way is to smooth the peaks and dips of the acoustic plant using the generalized complex smoothing technique suggested by Hatziantoniou and Mourjopoulos.[7] There are two alternative methods for implementing complex smoothing. The first method, uniform smoothing, is to calculate the impulse response using the inverse FFT of the frequency response. Then, apply a time-domain window to truncate and taper the impulse response, which in effect smoothes out the frequency response. Finally, recover the frequency response by FFT of the modified impulse response. Alternatively, a nonuniform smoothing method can also be used. This method performs smoothing directly in the frequency domain. The frequency response is circularly convolved with a window whose bandwidth increases with frequency. The choice of the window follows the psychoacoustics that the spectral resolution of human hearing increases with frequency. Therefore, the nonuniformly smoothed frequency response

$$H_{\mathrm{ns}}(m,k) = \sum_{i=0}^{N-1} H[(k-i) \bmod N] W_{\mathrm{sm}}(m,i), \tag{10}$$

where k, $0 \le k \le N-1$ is the frequency index and m is the smoothing index corresponding to the length of the smoothing window. The smoothing window $W_{\mathrm{sm}}(m,k)$ is given by

$$W_{\mathrm{sm}}(m,k) = \begin{cases} \dfrac{b-(b-1)\cos[(\pi/m)k]}{2b(m+1)-1}, & k = 0,1,\ldots,m \\ \dfrac{b-(b-1)\cos[(\pi/m)(k-N)]}{2b(m+1)-1}, & k = N-m, N-(m-1),\ldots,N-1 \\ 0, & k = m+1,\ldots,N-(m+1). \end{cases} \tag{11}$$

The integer, $m=m(k)$, can be considered as a bandwidth function by which a fractional octave or any other nonuniform frequency smoothing scheme can be implemented. The variable b determines the roll-off rate of the smoothing window. As a special case when $b=1$, the window reduces to a rectangular window.

C. Structures of inverse filters

There are a number of different ways to implement the inverse filters of CCS. The direct filtering method, the filter bank method, and the simple lowpass mixing method are three major filtering structures to discuss in this section.

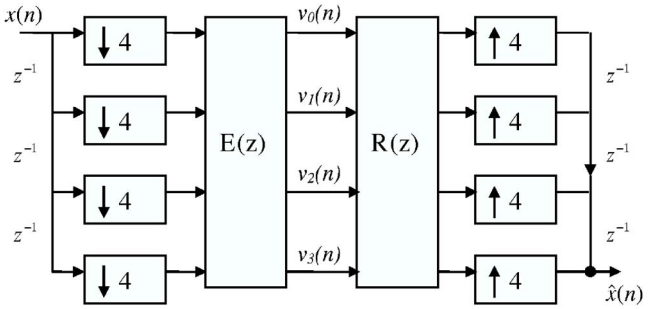

FIG. 2. The block diagram of a four-channel QMF bank using the polyphase representation.

1. Direct filtering method

In this structure, crosstalk cancellation is carried out by direct filtering using inverse filters. However, crosstalk cancellation can be demanded either for a full-band (200–24 kHz) performance or just a band-limited performance (200–6 kHz) in the design stage of inverse filters. The reason for the latter design is twofold. First, the sweet spot in which CCS is effective becomes impractically small at high frequencies. Second, a listener's head provides natural shadowing at high frequencies so that the need for cancellation becomes less important. The match equation appropriate for the band-limited design is written as[3]

$$\begin{bmatrix} \delta(n-m) & 0 \\ 0 & \delta(n-m) \end{bmatrix} = \begin{bmatrix} h_{11}(n) & h_{12}(n) \otimes f_{\mathrm{LP}}(n) \\ h_{21}(n) \otimes f_{\mathrm{LP}}(n) & h_{22}(n) \end{bmatrix} \otimes \begin{bmatrix} c_{11}(n) & c_{12}(n) \\ c_{21}(n) & c_{22}(n) \end{bmatrix}, \tag{12}$$

where $f_{\mathrm{LP}}(n)$ denotes the impulse response function of a lowpass filter. Thus, the inverse filters should in principle give rise to a flat response within the intended band after compensation.

2. Filter bank method

In the direct filtering approach, even if the inverse filters are deigned for band-limited performance, the filtering process is still carried out at a sampling rate of 48 kHz. To take advantage of the band-limited design, a subband filtering approach is exploited to simplify the computation. Specifically, a four-channel quadrature mirror filter (QMF) bank[13] is used to implement the CCS. For further enhancement of processing efficiency, the polyphase representation is employed to implement the QMF bank, as shown in Fig. 2. The block $\mathbf{E}(z)$ is the type 1 polyphase matrix for the analysis bank, and the block $\mathbf{R}(z)$ is the type 2 polyphase matrix for the synthesis bank. $v_i(n)$ represents the subband signal. The first subband signal is processed by the CCS and the other subband signals are simply delayed by the delay block $\mathbf{D}(z)$ and transmitted to the synthesis filter bank, as shown in Fig. 3.

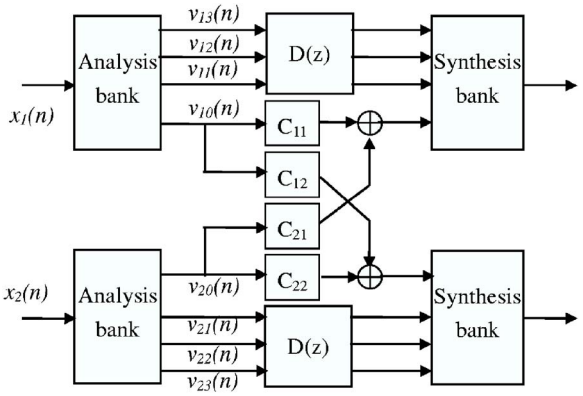

FIG. 3. Block diagram depicting the filter bank implementation of CCS.

3. Simple lowpass mixing method

For reference, a brief review of an alternative way of implementing the band-limited design originally proposed by Elliott *et al.* is also given (Fig. 4).[14] In this simple lowpass mixing approach, the input signal is lowpass filtered and down-sampled before sending to the CCS. Sufficient modeling delays must be inserted in the path. The CCS filters are adaptively updated by comparing the lowpass and delayed input and the lowpass plant out put at the control point (ears). Finally, the output of the CCS is up-sampled and remixed into the original full-band signal. The major difference between this method and the preceding filter bank method lies in the fact that the CCS-processed signal is mixed with the unprocessed full-band input in the simple mixing approach, while it is not the case in the filter bank method. This could have potential effect on the localization performance of spatializers.

D. Implemental issue

To facilitate the inverse filter design, the aforementioned smoothing techniques is employed to modify the impulse responses. On the other hand, the regularization parameters β and ε are selected to be 0.01 and 0.1 in the frequency-domain and time-domain deconvolutions, respectively, to limit the gain of the inverse filter to 10 dB maximum.

An objective index, channel separation, is employed to assess the cancellation performance

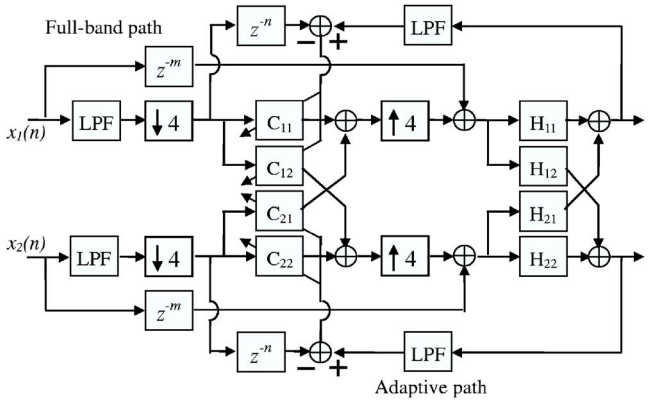

FIG. 4. Block diagram depicting the simple lowpass mixing implementation of CCS.

TABLE I. The test items used in the subjective evaluation.

Experiment 1	Test 1	Full-band frequency-domain CCS with uniform smoothing
	Test 2	Full-band time-domain CCS with uniform smoothing
Experiment 2	Test 1	Full-band conventional CCS with uniform smoothing
	Test 2	Full-band modified CCS-1 with uniform smoothing
	Test 3	Full-band modified CCS-2 with uniform smoothing
	Test 4	Commercial spatializer: DiMAGIC VX™ virtual sound imaging system
Experiment 3	Test 1	Full-band conventional CCS with uniform smoothing
	Test 2	Band-limited conventional CCS with uniform smoothing
	Test 3	Filter bank conventional CCS with uniform smoothing
	Test 4	Simple lowpass mixing conventional CCS with uniform smoothing
Experiment 4	Test 1	HRTF widening
	Test 2	Full-band conventional CCS with uniform smoothing
	Test 3	Full-band modified CCS-1 with uniform smoothing
	Test 4	HRTF+Full-band conventional CCS with uniform smoothing
	Test 5	HRTF+Full-band modified CCS-1 with uniform smoothing

$$S_{\mathrm{ep}}(j\Omega) = H_c(j\Omega)/H_i(j\Omega), \tag{13}$$

where $H_c(j\Omega)$ and $H_i(j\Omega)$ represent the contralateral (H_{12}, H_{21}) and the ipsilateral (H_{11}, H_{22}) frequency responses, respectively. According to the definition, a small (negative) value of channel separation indicates good cancellation performance.

III. DESIGN OF AUDIO SPATIALIZERS

A brief description of various approached based on HRTF and CCS will be given. For clarity, the experiments of audio spatializers were summarized in Table I.

A. HRTF

As mentioned previously, directional impression can be created by electronically synthesizing the HRTF in the desired angle. This is especially important in the case of mobile phones, where loudspeakers closely spaced. In this study, the HRTF database available in the website of the MIT media lab[1] was employed to "widen" the sound image. Each impulse response originally measured at a Knowles Electronic Mannequin for Acoustic Research (KEMAR) with a sampling frequency 44.1 kHz. HRTFs at the azimuth ±30° are implemented as 128-tapped finite impulse response (FIR) filters by which the audio input signals are filtered before sending to the loudspeakers. The processing can be written in matrix form as follows:

$$\begin{bmatrix} \hat{x}_1(n) \\ \hat{x}_2(n) \end{bmatrix} = \begin{bmatrix} h_{30\ \mathrm{ipsi}}(n) & h_{30\ \mathrm{contra}}(n) \\ h_{30\ \mathrm{contra}}(n) & h_{30\ \mathrm{ipsi}}(n) \end{bmatrix} \otimes \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}, \tag{14}$$

where $h_{30\ \mathrm{ipsi}}(n)$ and $h_{30\ \mathrm{contra}}(n)$ denote the ipsilateral and contralateral HRTFs, respectively, at the azimuths ±30°.

B. CCS

The objective of CCS is to minimize the effect of crosstalk. A generic inverse filter of a two-channel CCS can be a factored into the following expression:

$$\mathbf{C} = \frac{1}{1 - \mathrm{ITF}_1\mathrm{ITF}_2} \begin{bmatrix} 1/H_{11} & 0 \\ 0 & 1/H_{22} \end{bmatrix} \begin{bmatrix} 1 & -\mathrm{ITF}_2 \\ -\mathrm{ITF}_1 & 1 \end{bmatrix}, \tag{15}$$

where $\mathrm{ITF}_1 = H_{12}/H_{11}$, $\mathrm{ITF}_2 = H_{21}/H_{22}$ are interaural transfer functions, and the ipsilateral transfer functions H_{11}, H_{22} and the contralateral transfer functions H_{12}, H_{21} are defined as in Fig. 1. The earlier expression reveals the fact that the inverse filters attempt not only cancel the crosstalk with delays (the third term on the right hand side) but also equalize the ipsilateral response (the second term on the right hand side). The poles of the comb filter of the first term on the right hand side give the *ringing frequency*.[17]

The ipsilateral equalization (the second term) in the inverse filters may not be always desirable in practical application. For example, coloration problem may arise at around 10 kHz when the inverse filters strive to compensate the concha dip in the ipsilateral responses, which is largely independent of loudspeaker span. In addition, the other dips and roll-offs particularly at the very at the low and high frequencies in the ipsilateral responses further aggravate this situation. Consequently, an unnatural change of sound quality is often audible during reproduction due to over-compensating the ipsilateral responses. To address the problem, two modified techniques of CCS are suggested in the following.

1. The modified CCS-1

In this method, the diagonal terms of the matching model in the left hand side of Eq. (1) are replaced with delayed ipsilateral impulse responses

$$\begin{bmatrix} h_{11}(n-m) & \gamma \\ \gamma & h_{22}(n-m) \end{bmatrix} = \begin{bmatrix} h_{11}(n) & h_{12}(n) \\ h_{21}(n) & h_{22}(n) \end{bmatrix} \otimes \begin{bmatrix} c_{11}(n) & c_{12}(n) \\ c_{21}(n) & c_{22}(n) \end{bmatrix}, \quad (16)$$

where γ is a small constant, e.g., 0.0001 and m is the modeling delay. This in effect modifies the transfer functions of inverse filters in Eq. (15) into

$$C \approx \frac{1}{1-\mathrm{ITF}_1\mathrm{ITF}_2}\begin{bmatrix} 1 & -\mathrm{ITF}_2 \\ -\mathrm{ITF}_1 & 1 \end{bmatrix}. \quad (17)$$

The modified CCS makes no attempt to compensate the ipsilateral responses when canceling the crosstalk. It follows that the sound quality can be better preserved by using this method.

There is another potential benefit in the use of this method. Assume that two speaker responses are displaced by a factor S. Neglecting the parameter γ, the z-domain version of Eq. (16) can be written as

$$\begin{bmatrix} z^{-m}\tilde{H}_{11}(z)S & 0 \\ 0 & z^{-m}\tilde{H}_{22}(z)S \end{bmatrix} \approx \begin{bmatrix} \tilde{H}_{11}(z)S & \tilde{H}_{12}(z)S \\ \tilde{H}_{21}(z)S & \tilde{H}_{22}(z)S \end{bmatrix}\begin{bmatrix} C_{11}(z) & C_{12}(z) \\ C_{21}(z) & C_{22}(z) \end{bmatrix}, \quad (18)$$

where $\tilde{H}_i$ represents the transfer function without loudspeaker responses. Thus, the factor S cancels out on both sides. The implication of this is that the CCS is loudspeaker independent as long as the characteristics of two loudspeakers are well matched. This could be a desirable property in practical applications in that a CCS designed off-line is applicable to all systems with different loudspeaker characteristics.

2. The modified CCS-2

Along the same line, another modified CCS is developed to underplay the equalization of ipsilateral response during cancellation of crosstalk. In this approach, the ipsilateral inverse filters are assigned to be a delayed discrete delta function, i.e., $c_{11}=c_{22}=\delta(n-m)$ such that the sound quality can be preserved because of the direct transmission of ipsilateral paths. In this setting, the match equation should be modified into

$$\begin{bmatrix} d_L(n) & 0 \\ 0 & d_R(n) \end{bmatrix} = \begin{bmatrix} h_{11}(n) & h_{12}(n) \\ h_{21}(n) & h_{22}(n) \end{bmatrix} \otimes \begin{bmatrix} \delta(n-m) & c_{12}(n) \\ c_{21}(n) & \delta(n-m) \end{bmatrix}, \quad (19)$$

where the diagonal terms d_L and d_R are the resulting ipsilateral responses. Expanding this equation only for the off-diagonal terms leads to two equations

$$-(h_{12}(n) \otimes \delta(n-m)) = -h_{12}(n-m) = h_{11}(n) \otimes c_{12}(n), \quad (20)$$

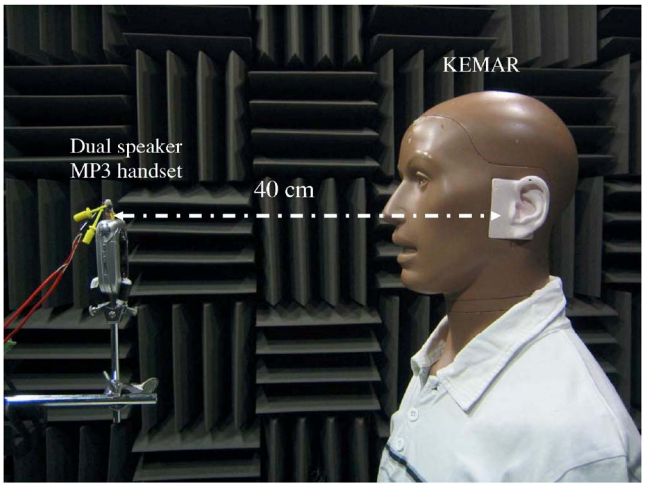

FIG. 5. Experimental arrangement for the dual speaker handset with a dummy head system inside an anechoic chamber.

$$-(h_{21}(n) \otimes \delta(n-m)) = -h_{21}(n-m) = h_{22}(n) \otimes c_{12}(n). \quad (21)$$

The contralateral inverse filters can be obtained by solving this inverse problem. By the same token, it can be shown that this modified CCS is also loudspeaker independent. However, this approach would possibly lead to poor bass response because the crosstalk canceller will no longer have the factor $1/(1-\mathrm{ITF}^2)$, which is essentially a bass boost.

IV. EXPERIMENTAL INVESTIGATIONS

A. Experimental arrangement

The experiments were conducted by using a dummy head system (KEMAR) inside a 4 m×4 m×3 m anechoic chamber, as shown in Fig. 5. An MP3 handset equipped with dual loudspeakers is mounted on a stand. The distance between the handset, and the dummy head is 40 cm. Binaural transfer functions from the loudspeakers to the microphone embedded in the dummy head's ears were measured by using a spectrum analyzer. The algorithms were implemented on the platform of a fixed-point DSP, ADI BF-533, operating at 48 kHz. The inverse filters were realized as 128-tapped FIR filters in the experiments.

B. Objective experiment

For simplicity, symmetrical acoustic plant is assumed. The head-related impulse responses measured by using the dummy head is shown in Fig. 6. The complex smoothing is applied prior to the design of CCS. In this regard, the CCS will prove more robust against misalignment of the listener's head than that designed for unsmoothed frequency responses.[10,18] Figure 7 shows frequency responses obtained using uniform smoothing and nonuniform smoothing. It can be seen that the frequency responses are effectively smoothed by both methods. However, an informal subjective test has indicated that the difference between the two smoothing techniques is hardly detectible. The uniform smoothing method, therefore, is used exclusively in the following experiments.

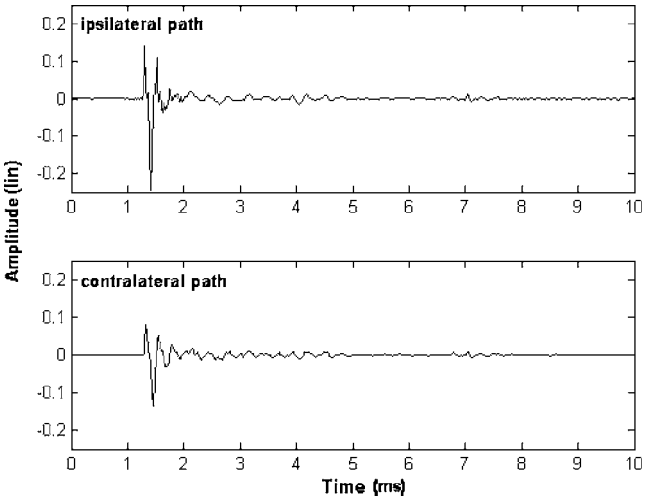

FIG. 6. Head-related impulse responses measured by using the dummy head system.

Another issue concerning the CCS design is a modeling delay that is necessary to ensure the causality of inverse-filters. This is of fundamental importance whether the frequency-domain method or the time-domain method is used. A simple experiment was conducted to examine the effect of different modeling delays on a 128-tapped filter and a 512-tapped taps filter obtained using the time-domain method. Average channel separation (Ave-Sep, dB) between 200 and 20 kHz is calculated to assess the cancellation performance. The result summarized in Table II reveals that the optimal modeling delay is approximately half of the length of the inverse filter.

The length of the inverse filter also affects the performance of CCS. The performance of inverse filters with different length of inverse filter is compared in Table III. As expected, the performance of CCS improves as the filter length is increased for both deconvolution methods. However, it is worth noting that the time-domain method outperforms the frequency-domain method for short filter length such as 128 taps. The frequency-domain method performs well only when a long filter is used. Another drawback of the frequency-domain method can be clearly seen by plotting the magnitudes of the equalized time responses on the dB scale, as suggested by Fielder.[19] In Fig. 8(a), preringing artifacts are visible (at 1–3 ms) in the equalized time responses when the frequency-domain method is used, while no such artifacts are found in the result of the time-domain method in Fig. 8(b).

Next, a useful variation of inverse filter design to enhance CCS performance is examined. Figure 9(a) shows the experimental results of the unprocessed and the processed frequency responses with the conventional CCS. While the flat spectrum is attained as expected in the compensated ipsilateral response, the contralateral response is not totally eliminated but amplified at the frequencies above 10 kHz. This incurs some audible coloration at high frequencies. To overcome the problem, the aforementioned modified approaches were employed to suppress the crosstalk while preserving the ipsilateral response. Figures 9(b) and 9(c) refer to the implementation of the modified CCS-1 and CCS-2, respectively. It is observed that not only the ipsilateral response remains largely unchanged but also the contralateral response is effectively attenuated without undesired amplification in high frequencies. To explore further the modified CCS, the time responses of the inverse filters of the modified methods are compared with those obtained using the conventional identity matching model. Figure 10(a) refers to the implementation of the conventional CCS. Figures 10(b) and

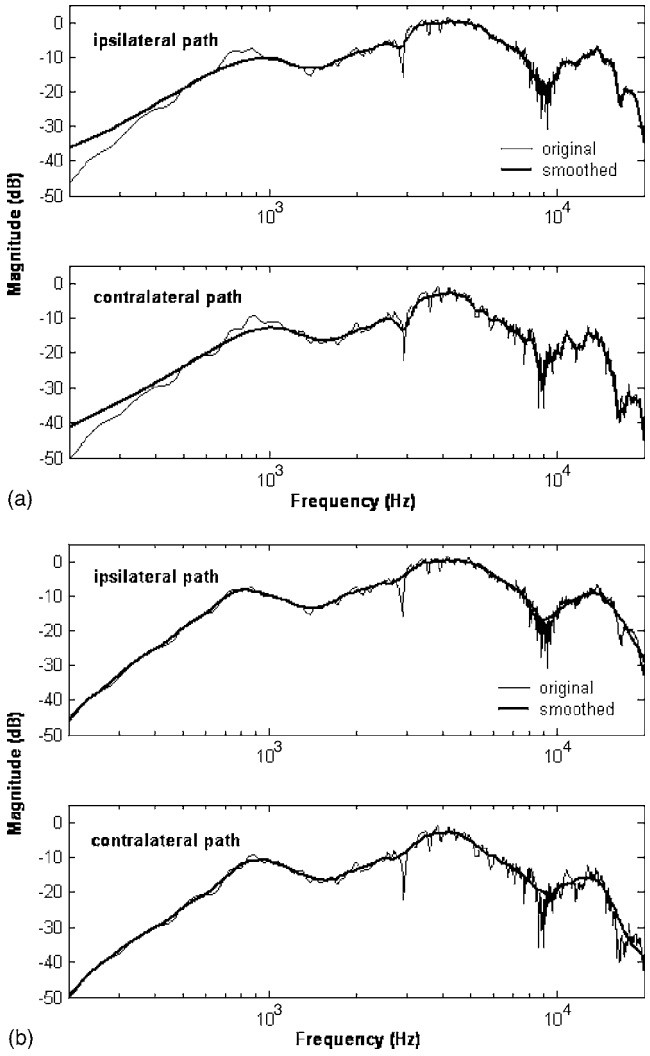

FIG. 7. Comparison between original and the complex smoothed magnitude spectrum. The thick line represents the complex smoothed magnitude response spectrum. (a) Result obtained using uniform smoothing. (b) Result obtained using nonuniform smoothing.

TABLE II. The average separation obtained using the time-domain method with different delays.

Filter length (N_c): 128 taps		Filter length (N_c): 512 taps	
Delay (m)	Average separation (dB)	Delay (m)	Average separation (dB)
16	−20.583	32	−20.799
32	−20.717	128	−21.592
48	−20.833	256	−21.701
64	−21.050	288	−21.706
80	−21.007	320	−21.692
96	−20.282	448	−20.771

TABLE III. The average separation obtained using inverse filtering with different filter length.

Filter length (N_c)	Average separation (dB)	
	Frequency domain	Time domain
128	−18.203	−21.050
256	−18.361	−21.608
512	−21.535	−21.705
1024	−22.329	−21.760
2048	−22.375	−21.870

10(c) refer to the implementation of the modified CCS-1 and CCS-2, respectively. The impulse responses of inverse filters designed using the modified methods are significantly shorter than those of the conventional method. This computational saving is a benefit for real-time implementation.

The inverse filters were implemented by using the band-limited design as detailed in the preceding section. It can be seen in the experimental result of Fig. 11 that the CCS maintains wideband equalization of the ipsilateral response to result in a flat spectrum, while the cancellation of crosstalk is only attained in low frequency range with some unwanted amplification in the high frequency range. Cancellation performance is confined in low frequency range as it should be for the filter bank method and the simple lowpass mixing method since they are essentially band-limit designs.

C. Subjective experiment

In order to assess the perceptual performance of the spatializers, subjective listening tests were conducted according to the double-blind triple stimulus with hidden reference method suggested in the standard ITU-R BS. 1116-1.[20] The listening tests were carried out inside the anechoic chamber. The program material consists of various instruments with significant dynamic variations between the two stereo channels. Both timbre-related and space-related qualities are considered. The loudness of each reproduced signal was adjusted with equal power. Nine subjective indices employed in the subjective tests are summarized as follows:

(1) Fullness: Dominance of low-frequency sound;
(2) Brightness: Dominance of high-frequency sound;
(3) Noise and distortion: Any extraneous disturbances to the signal are considered as noise. Effect on the signal that produces new sounds or timbre change is considered as distortion;
(4) Width of stage: Perceived angular width of extreme left to extreme right edges of the stage;
(5) Depth perception: Ability to hear that performers are appropriately localized from the front to the rear of the sound stage;
(6) Spaciousness: Perceived quality of listening within a reverberant environment. The sound is perceived as open, not constrained to the locations of the loudspeakers. The perception is an important part of the "you are there" sensation;
(7) Localization: Determination by a subject of the apparent direction or distance, or both, of a sound source;

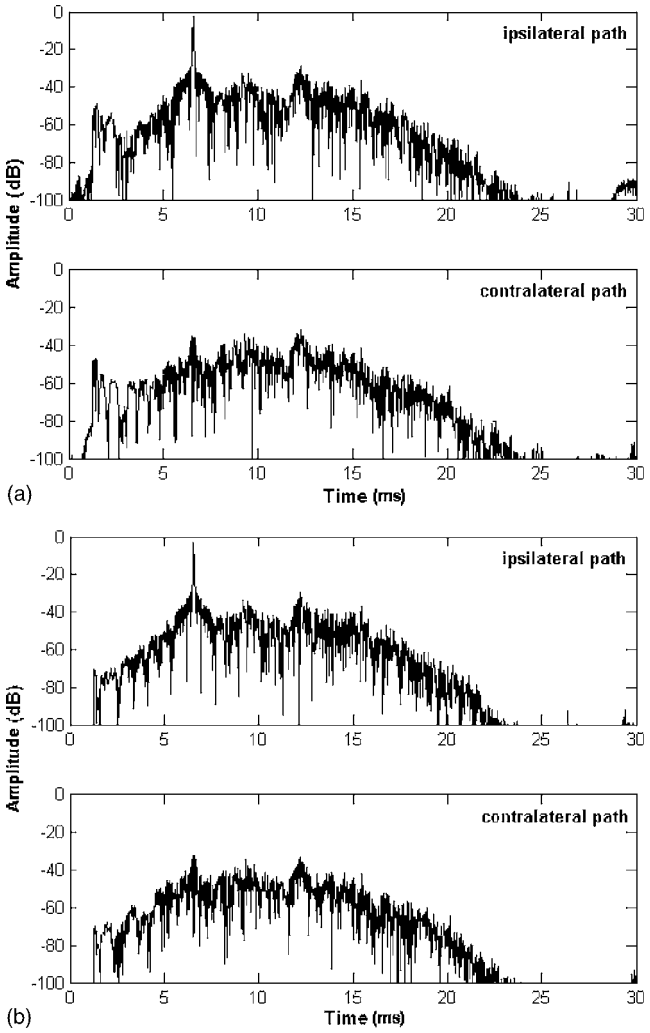

FIG. 8. Equalized time responses plotted on the dB scale. (a) Frequency-domain method. (b) Time-domain method.

(8) Robustness: Stability of performance with normal listener movements and listening locations. This index is assessed by 5 and 10 cm lateral movement of listener's head, and calculating the average grade; and
(9) Fidelity: The clarity of the reproduced signals.

Twenty experienced subjects participating in the tests were instructed with definition of the preceding subjective indices and the procedure before the listening tests. The subjects were asked to respond after listening in a questionnaire, with the aid of a set of subjective indices placed on a scale from −4 to 4. Positive, zero, and negative scores indicate perceptually improvement, no difference, and degradation, respectively, of the signals after processed by the spatializers. In order to justify the statistical significance, the scores were further processed by using the MANOVA.[15] Cases with significance levels below 0.05 indicate that statistically significant difference exists among methods. The experiments were summarized in Table I.

The first listening test was carried to compare the frequency-domain and the time-domain methods. The total grades are plotted in Fig. 12. The vertical bars denote 0.95 confidence intervals. The small significance level (s

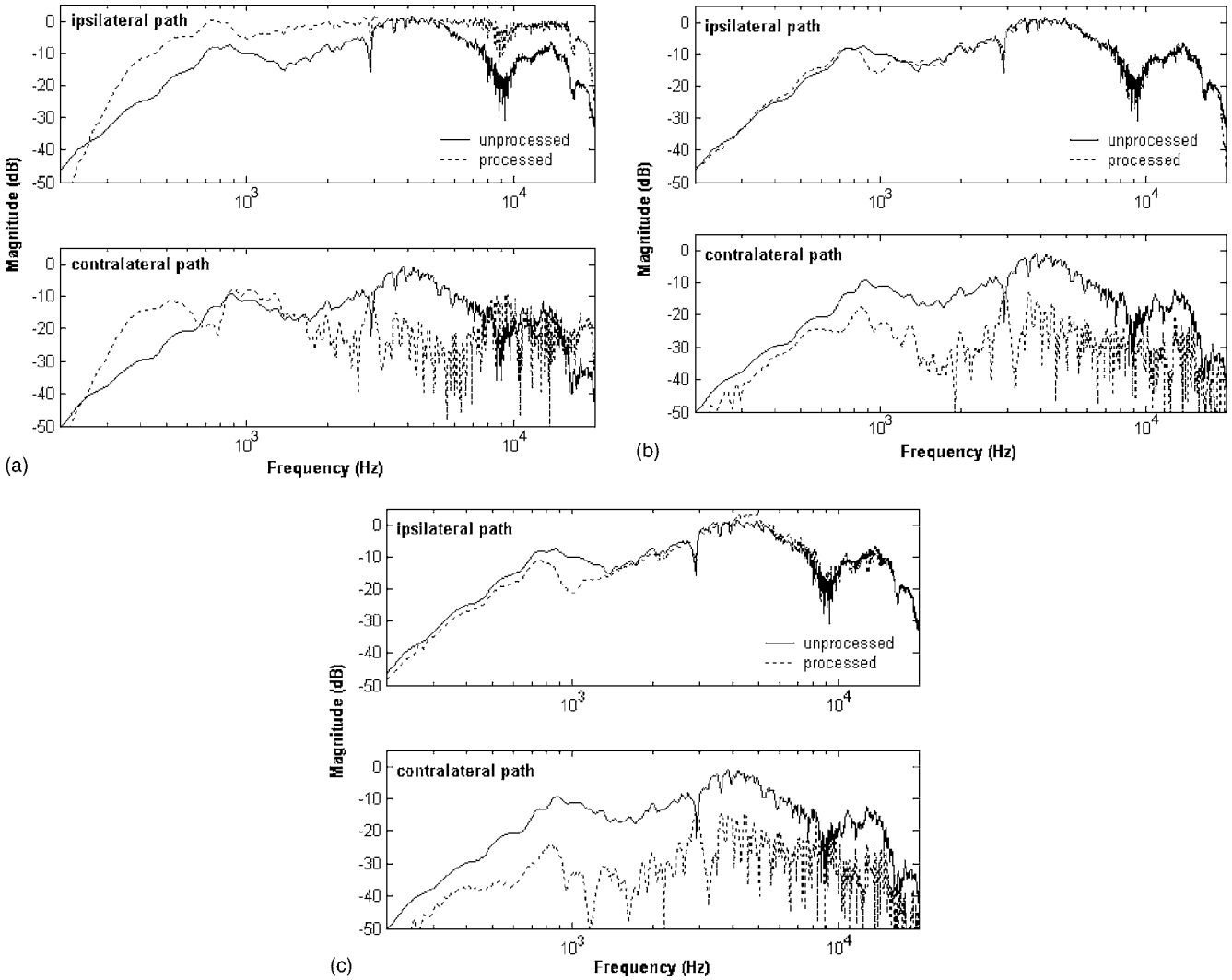

FIG. 9. Comparison between the unprocessed and the processed frequency responses. (a) The conventional CCS. (b) The modified CCS-1. (c) The modified CCS-2.

=0.041523) in the MANOVA output indicates that the difference among the methods is statistically significant. In particular, the time-domain method seemed to significantly outperform the frequency-domain method for inverse filters of this length (128 taps), which is in agreement with the observation in the preceding objective tests.

Next, the second experiment is performed to compare the modified CCS methods and a commercial spatializer[21] which is used in this experiment as the benchmark. The results shown in Fig. 13 revealed that the modified method-1 received the highest score among all approaches with strong statistical significance (s=0.000001). The modified method-1 is particularly advantageous when sound quality is used as the performance index in addition to the cancellation performance.

In the third listening test, different structures of CCS implementation are compared. The total grades are summarized in Fig. 14. The MANOVA output reveals that significant difference in performance (s=0.019207) does exist among the methods. The direct filtering method has attained the highest grade, while the simple lowpass mixing method received the lowest grade. In the direct filtering approach, there is no significant difference between the full-band and the band-limited designs. It is worth noting that the filter bank approach and the simple lowpass mixing approach did not attain the grades as high as two other direct filtering approaches. Possible explanations for this are that the crossovers in the filter bank are not adequately handled in the filter bank methods, and portion of the low-frequency signal is contaminated by crosstalk in the simple lowpass mixing method.

In the fourth listening test, various audio spatializers utilizing the HRTF, the conventional CCS method, the modified CCS method-1, and their combinations are compared. The total grades are summarized in Fig. 15. The MANOVA output reveals that significant difference in performance (s =0.000001) exists among the methods. It is observed from the result that the HRTF approach receives the lowest grade. The "widening" effect provided by the HRTF solely is obviously insufficient to spatialize the sound image due to the severe crosstalk between the closely spaced loudspeakers. In contrast to the HRTF approach, there is a leap in performance when the CCS comes into play. In particular, the spatializer combining the HRTF and the conventional CCS

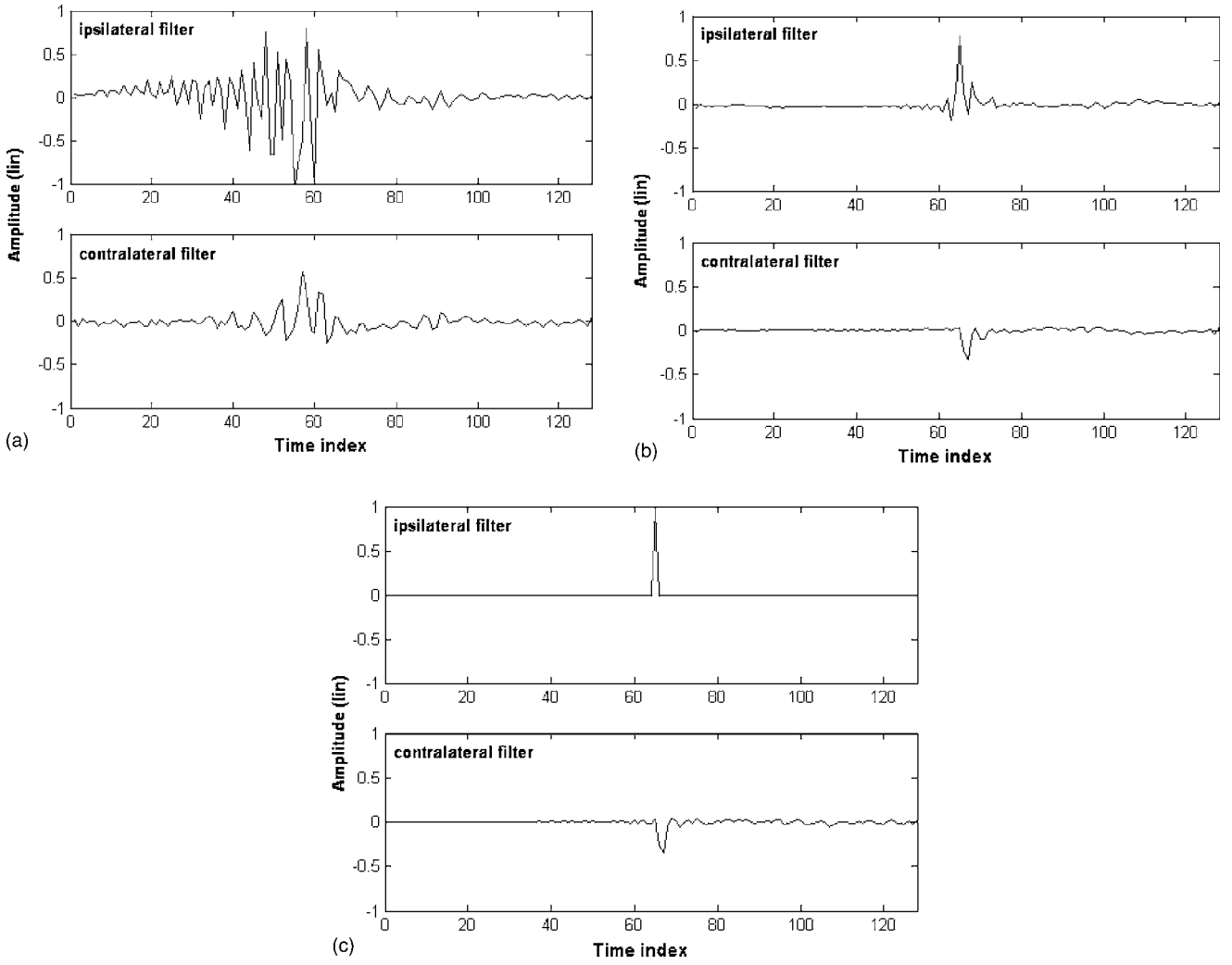

FIG. 10. Impulse responses of the inverse filters. (a) The conventional CCS method. (b) The modified CCS-1. (c) The modified CCS-2.

method has achieved the highest grade in both spatializing performance and sound quality. Surprisingly, when the modified CCS method is used in combination with the HRTF, there is a sudden drop in performance. It is suspected that double HRTF filtering effect may have contributed to this result. That is, while the sound quality has already been preserved by plugging the HRTF in the matching model for the modified CCS, the additional HRTF filtering becomes superfluous and may adversely affect the sound quality of the processed signal.

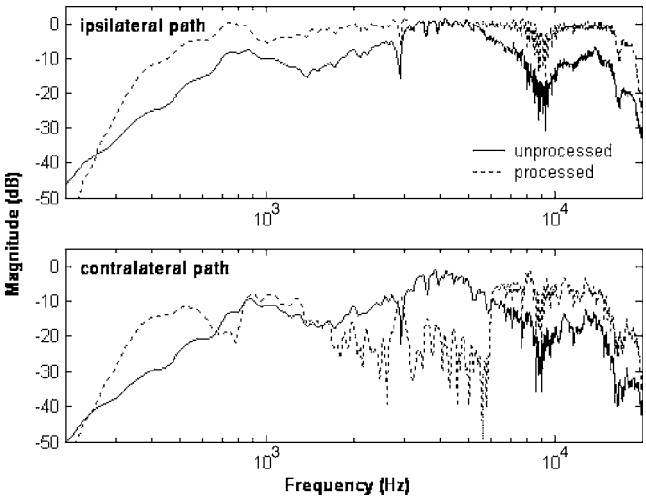

FIG. 11. Comparison between the unprocessed frequency response and that processed by using the band-limited CCS.

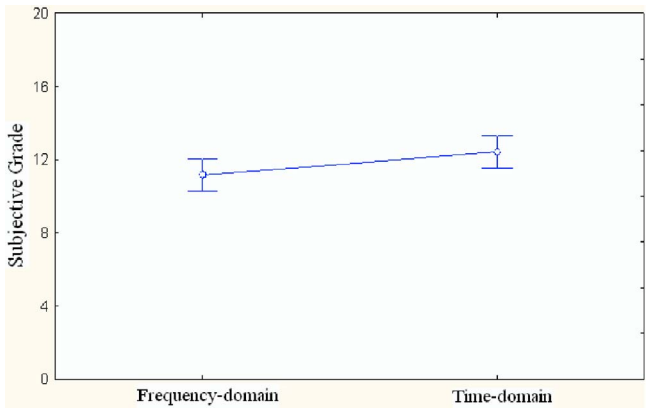

FIG. 12. Total grades summarized for the first listening test in which the frequency-domain and the time-domain deconvolution methods are compared. The significance level, $s=0.041523$, in the MANOVA output.

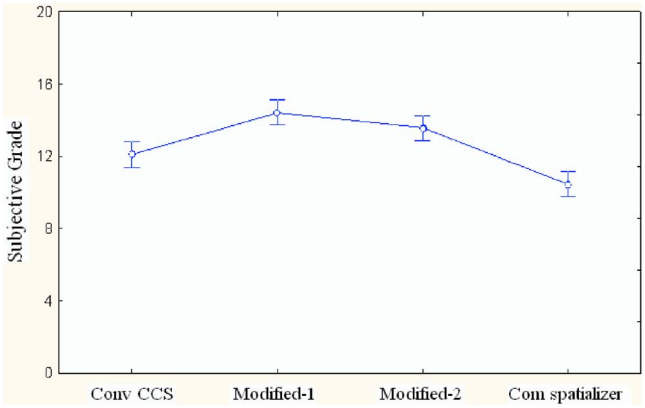

FIG. 13. Total grades summarized for the second listening test in which various CCS approaches including a commercial spatializer are compared. (Conv CCS: conventional CCS with identity matching, modified-1: modified CCS-1, modified-2: modified CCS-2, com spatializer: DiMAGIC VX™ virtual sound imaging system.) The significance level, $s=0.000001$, in the MANOVA output.

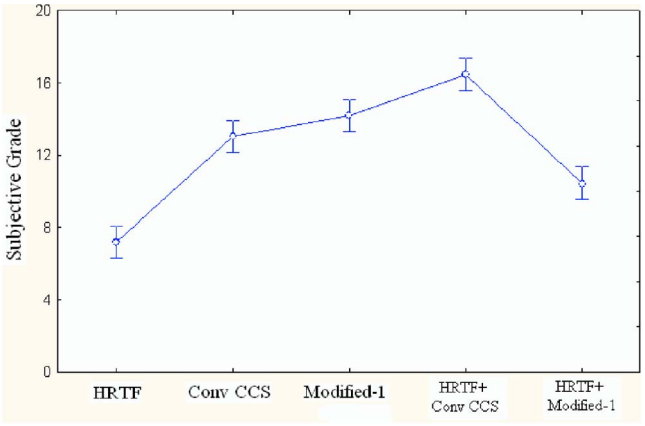

FIG. 15. Total grades summarized for the fourth listening test in which various audio spatializers utilizing the HRTF, the conventional CCS method, the modified CCS-1, and their combinations are compared. (HRTF: HRTF widening, conv CCS: conventional CCS, modified-1: modified CCS-1, HRTF+conv CCS: HRTF combined with conventional CCS, HRTF +modified-1: HRTF combined with modified CCS-1.) The significance level, $s=0.000001$, in the MANOVA output.

V. CONCLUSIONS

A comprehensive study has been undertaken to compare various implementation approaches of audio spatializer for handsets fitted with two closely spaced loudspeakers. The HRTF and the CCS techniques were exploited to implement the audio spatializer. Two deconvolution methods were applied to calculate the inverse filters for the CCS design. Objective and subjective experiments reveal that the time domain approach is superior to the frequency-domain approach when the length of inverse filter is short. An additional benefit of the time-domain method is that it is less liable to preringing artifact that frequently appears in the frequency-domain method.

Different structures of CCS were examined in this study. The experimental results indicate that the direct filtering approaches outperform the filter bank method and the simple lowpass mixing method. In addition, two modified CCS techniques were proposed in the present paper. Unlike the conventional method that tends to over-compensate the ipsilateral responses, the modified methods are capable of delivering better spaciousness without compromising on sound quality. Two additional features of the modified CCS which are attractive in practical application lie in its shorter impulse responses of the inverse filters and the loudspeaker-independent property.

Listening tests were also carried out to compare various ways of implementing a spatializer based on HRTF, CCS, and their combination. The experimental results suggest that the widening effect provided by the HRTF solely is insufficient to spatialize the sound image due to the severe crosstalk between the closely spaced loudspeakers. In contrast to the HRTF approach, there is a leap in performance when the CCS is used. In particular, the spatializer combining the HRTF and the conventional CCS method has achieved the best performance in both spatializing performance and sound quality.

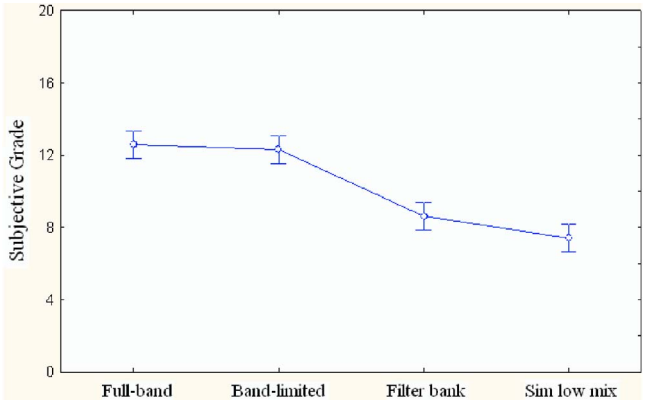

FIG. 14. Total grades summarized for the third listening test in which different structures of CCS implementation are compared. (full-band: full-band CCS, band-limited: band-limited, filter bank: filter bank CCS, sim low mix: simple lowpass mixing CCS.) The significance level, $s=0.019207$, in the MANOVA output.

ACKNOWLEDGMENT

The work was supported by the National Science Council in Taiwan, Republic of China, under the Project No. NSC94-2212-E-009-019.

[1]B. Gardner and K. Martin, "HRTF measurements of KEMAR dummy-head microphone," MIT Media Lab, 1994; http://sound.media.mit.edu/KEMAR.html. Last accessed 11/10/06.
[2]A. Sibbald, "Transaural acoustical crosstalk cancellation," Sensaura White Paper, 1999, http://www.sensaura.co.uk. Last accessed 11/10/06.
[3]W. G. Gardner, *3-D Audio Using Loudspeakers* (Kluwer Academic, Dordrecht, 1998).
[4]O. Kirkeby, P. A. Nelson, and H. Hamada, "Fast deconvolution of multichannel systems using regularization," IEEE Trans. Speech Audio Process. **6**, 189–195 (1998).
[5]O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," J. Audio Eng. Soc. **47**, 583–595 (1999).
[6]J. F. Claerbout, *Earth Soundings Analysis: Processing Versus Inversion (PVI)*, 1992; http://sep.stanford.edu/sep/prof/toc_html/index.html. Last accessed 11/10/06.
[7]P. D. Hatziantoniou and J. N. Mourjopoulos, "Generalized fractional-octave smoothing of audio and acoustic responses," J. Audio Eng. Soc. **48**, 259–280 (2000).
[8]S. G. Norcross, G. A. Soulodre, and M. C. Lavoie, "Subjective investigations of inverse filtering," J. Audio Eng. Soc. **52**, 1003–1028 (2004).

[9]S. Neely and J. B. Allen, "Invertibility of a room impulse response," J. Acoust. Soc. Am. **66**, 165–169 (1979).
[10]P. M. Clarkson, J. Mourjopoulos, and J. K. Hammond, "Spectral, phase, and transient equalization for audio systems," J. Audio Eng. Soc. **33**, 127–132 (1985).
[11]O. Kirkeby, P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Local sound field reproduction using digital signal processing," J. Acoust. Soc. Am. **100**, 1584–1593 (1996).
[12]J. H. Wang and C. S. Pai, "Subjective and objective verifications of the inverse functions of binaural room impulse response," Appl. Acoust. **64**, 1141–1158 (2003).
[13]P. P. Vaidyanathan, *Multirate Systems and Filter Banks* (Prentice-Hall, Englewood Cliffs, NJ, 1993).
[14]S. J. Elliott, P. A. Nelson, and I. M. Stothers, "Sound reproduction systems," U.S. Patent No. 5,727,066 (1998).
[15]G. Keppel and S. Zedeck, *Data Analysis for Research Designs* (W. H. Freeman, New York, 1989).
[16]H. Hamada, "Construction of orthostereophonic system for the purposes of quasiinsitu recording and reproduction," J. Acoust. Soc. Jpn. **39**, 337–348 (1983).
[17]J. Rose, P. A. Nelson, B. Rafaely, and T. Takeuchi, "Sweet spot size of virtual acoustic imaging systems at asymmetric listener locations," J. Acoust. Soc. Am. **112**(5), 1992–2002 (2002).
[18]S. Salamouris, K. Politopoulos, V. Tsakiris, and J. Mourjopoulos, "Digital system for loudspeaker and room equalization," J. Audio Eng. Soc. **43**, 396 (1995).
[19]L. D. Fielder, "Analysis of traditional and reverberation-reducing method of room equalization," J. Audio Eng. Soc. **51**, 3–26 (2003).
[20]ITU-R BS. 1116, "Methods for the subjective assessment of small impairments in audio system including multichannel sound systems," Geneva, Switzerland, 1994.
[21]DiMAGIC, "DiMAGIC VX™ virtual sound imaging system," White Paper, 2000; http://www.dimagic.com/pdf/DiMAGIC_Virtualizer_X_White_paper.pdf. Last accessed 11/10/06.