Upmixing and Downmixing Two-channel Stereo Audio for Consumer Electronics

Mingsian R. Bai and Geng-Yu Shih

**Abstract — *In this comprehensive study, algorithms for upmixing, downmixing, and joint up/downmixing are examined and compared. Five upmixing algorithms based on signal decorrelation and reverberation are employed to convert two-channel stereo signals to five-channel signals. For downmixing, methods ranging from mixing with simple gain adjustment to more sophisticated Head Related Transfer Function (HRTF) filtering and Crosstalk Cancellation System (CCS) are utilized to downmix the center channel and the surround channels into the available two frontal loudspeakers. For situations where only two-channel content and loudspeakers are available, a number of up/down mixing schemes are used to simulate a virtual surround environment. Emphasis of comparison is placed on two consumer electronic products: a 5.1 home theater system and a dual-loudspeaker MP3 handset. The effect of loudspeaker spacing on rendering performance is examined. Listening tests are conducted to compare the processing methods in terms of three levels of subjective indices. The results are processed by using the Multi-Analysis Of VAriance (MANOVA) to justify the statistical significance, followed by a multiple regression analysis to correlate the auditory preference with various timbral and spatial attributes*[1].

Index Terms —Virtual surround, upmixing, downmixing, subjective listening test.

I. INTRODUCTION

With the growing proliferation of multichannel audio in consumer electronics, there are situations where the number of channels of either the audio content or the reproducing loudspeakers is limited and it is necessary to upmix or downmix channels to improve the listening experience. Yet, another situation where combined upmixing and dwnmixing is necessary is the newly emerging third-generation (3G) handsets fitted with dual loudspeakers. This paper is aimed at a comprehensive study that systematically compares a variety of upmixing and downmixing strategies.

In the upmixing process, surround channels can be created from the two-channel stereo inputs by using two different approaches. One approach uses the 'decorrelated' part of the stereo inputs as the surround channels, whereas the other approach produces the surround channels by simulating the reverberant sound field in the background. In this paper, five techniques including a passive surround decoder method [1], a Least-Mean-Square (LMS)-based method [2], an adaptive panning method [3], a Principal Component Analysis (PCA)-based method [4], and an artificial reverberator-based method [5] are presented.

Contrary to the upmixing process, the downmixing process is employed to produce a decreased number of channels due to practical reasons such as availability of loudspeakers. Without loss of generality, this study is focused on remixing 5.1 audio inputs into two-channel signals. Downmixing can be accomplished by simple mixing or more complex filtering by Head Related Transfer Functions (HRTFs), as in the Sound Retrieval System (SRS) 3D stereo sound system [6]. An HRTF is a mathematical model representing the propagation process from a sound source to the human ears [7].

Another problem well known in reproduction using loudspeakers is that the crosstalk of the contralateral paths from the loudspeakers to the listener's ears can adversely affect source localization. A solution to this problem is to minimize crosstalk using a Crosstalk Cancellation System (CCS) derived from inverse filter design. In this paper, downmixing techniques, ranging from mixing with simple gain adjustment [8] to more sophisticated HRTF filtering [9], and CCS-based processing [10], shall be examined.

One last situation that may call for the combined use of upmixing and downmixing is when the audio inputs and the reproducing loudspeakers are both of the two-channel stereo configurations. The interactions between upmixing and downmixing are also investigated in the paper.

Subjective tests are carried out to assess the performance of each processing method. The experimental results are processed by using the multi-analysis of variance (MANOVA) to justify the statistical significance. In addition, a multiple regression model is employed to correlate global auditory preference with low-level attributes. In light of these comprehensive tests, it is hoped that viable upmixing and downmixing techniques can be found to cater for practical multichannel audio reproduction.

II. UPMIXING ALGORITHMS

In this section, five upmixing algorithms are introduced. These methods differ in how to derive the additional channels from the two stereo channels. The general architecture of the direct-ambient upmixing approach for creating the additional channels is shown in Fig. 1. Common to all methods, the Center (C) channel is created by filtering the correlated input using a 128-tapped FIR bandpass filter with cut-off frequencies 100 Hz and 4 kHz to emphasize voice and dialog. The Rear Left (RL)

[1] The work was supported by the National Science Council in Taiwan, Republic of China, under the project number NSC94-2212-E-009-019.

The authors are with the Department of Mechanical Engineering, the National Chiao-Tung University, 1001 Ta-Hsueh Road, Hsin-Chu 300, Taiwan. (e-mail: msbai@mail.nctu.edu.tw).

Contributed Paper
Manuscript received May 30, 2007

and the Rear Right (RR) channels are intended to provide ambience and envelopment in the background, where a 15 ms delay is added to the rear channels to fulfill the precedence effect. High-frequency absorption is simulated by filtering the rear channels with a 7 kHz cut-off, 128-tapped FIR lowpass filter. In addition, the rear channels are 180-degree out-of-phase with one another, which can increase spaciousness of the ambient field [11]. The Low Frequency Enhancement (LFE) channel is derived from the center channel with a 128-tapped FIR lowpass filter to retain the signals below 120 Hz.

Five upmixing techniques are presented as follows.

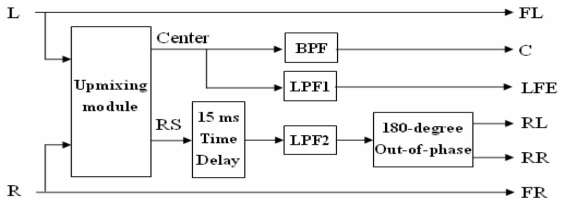


Fig. 1. General architecture of direct-ambient upmixing technique for the center, the LFE, and the surround channels.

A. The Passive Surround Decoder

The first approach employed in this study follows from an early passive version of the Dolby Surround Decoder [1], as shown in Fig. 2. The center channel is the average of the original stereo channels, whereas the rear surround signals are the difference of the stereo channels. That is,

$$\text{Center} = (L+R)\big/\sqrt{2} \tag{1}$$

$$\text{Surround} = (L-R)\big/\sqrt{2} \tag{2}$$

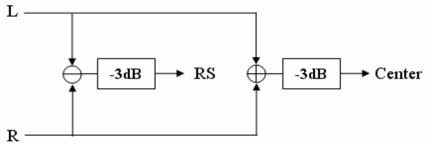


Fig. 2. The block diagram of the passive surround decoder.

B. The LMS-based Method

The approach to be described is based on a decorrelation technique using the LMS algorithm [2]. The basic idea of the LMS filter is shown in Fig. 3, where $d(n)$ is the desired signal, $\mathbf{x}(n)$ is the input of an FIR filter, $\mathbf{w}(n)$ is the coefficient vector of a 16-tapped FIR filter, $y(n)$ is the output of an FIR filter and $e(n)$ is the error signal. The LMS algorithm consists of three important equations:

$$y(n) = \mathbf{x}^T(n)\mathbf{w}(n) = \mathbf{w}^T(n)\mathbf{x}(n) \tag{3}$$

$$e(n) = d(n) - y(n) = d(n) - \mathbf{w}^T(n)\mathbf{x}(n) \tag{4}$$

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu e(n)\mathbf{x}(n), \tag{5}$$

where μ is a constant step size that dictates the convergence behavior. The left channel and the right channel of the original stereo channels are taken as the desired signal and the input of the FIR filter, respectively. The output $y(n)$ representing the correlated part is used as the center channel, whereas the error $e(n)$ representing the uncorrelated part is used to produce two surround channels in upmixing. A special implementation of LMS algorithm called Normalized LMS (NLMS) takes into account the variation in the input signals and selects a step size normalized with the input power. The weight update equation is modified into

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{\tilde{\mu}}{\mathbf{x}^T(n)\mathbf{x}(n) + \psi} e(n)\mathbf{x}(n), \tag{6}$$

where $\tilde{\mu}$ and ψ are positive constants.

In some occasions, the adaptive algorithm diverges due to low correlation level between the signals of the front channels when dominated with diffuse components. To deal with it, a correlation-based method can be used, where the step size is selected according to a modified correlation coefficient [12]:

$$\rho(n) = \left|\sum_{i=0}^{M-1} x(n-i)d(n-i)\right| \Big/ \sum_{i=0}^{M-1} \left|x(n-i)d(n-i)\right|, \tag{7}$$

where M is the length of the estimation window.

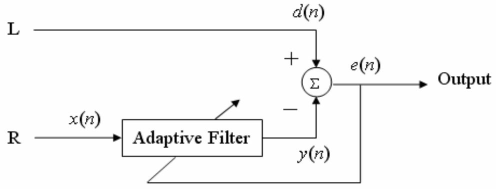


Fig. 3. The adaptive LMS algorithm with the left channel as the desired signal and the right channel as the input of a FIR filter.

C. The Adaptive Panning Method

Another method to be considered in this section is proposed by Irwan and Aarts [3]. Let $x_L(n)$ and $x_R(n)$ be the stereo signals, as shown in Fig. 4. The dominant signal $y(n)$ and the remaining signal $q(n)$ are generated by simple gain adjustment of the input signals.

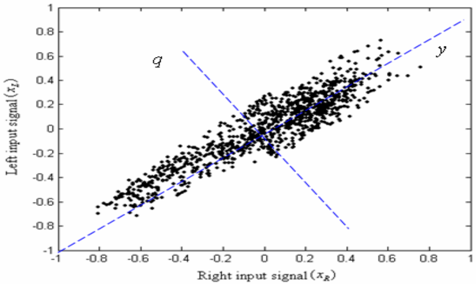


Fig. 4. The plot of original stereo signals. Dashed lines represent new coordinate system based on both dominant signal y and remaining signal q, forming the direction of principal axes.

To maximize the energy of $y(n)$, two scalar panning weights, $w_L(n)$ and $w_R(n)$, corresponding to the left and right channels are determined by using the LMS algorithm with $y(n\text{-}1)$ being the input.

$$\begin{aligned} w_L(n) &= w_L(n-1) + \mu y(n-1)[x_L(n-1) - w_L(n-1)y(n-1)] \\ w_R(n) &= w_R(n-1) + \mu y(n-1)[x_R(n-1) - w_R(n-1)y(n-1)] \end{aligned} \tag{8}$$

Two signals, $y(n)$ and $q(n)$, are then obtained by panning the stereo signals $x_L(n)$ and $x_R(n)$ with the weights found above.

$$y(n) = w_L(n)x_L(n) + w_R(n)x_R(n) \tag{9}$$

$$q(n) = w_R(n)x_L(n) - w_L(n)x_R(n) \tag{10}$$

Consequently, the dominant signal $y(n)$ and the remaining signal $q(n)$, corresponding to the correlated and the decorrelated channel, respectively, are used as the center and the surround channels in upmixing. The weights $w_L(n)$ and $w_R(n)$ generally fluctuate around 0.7, as shown in Fig. 5.

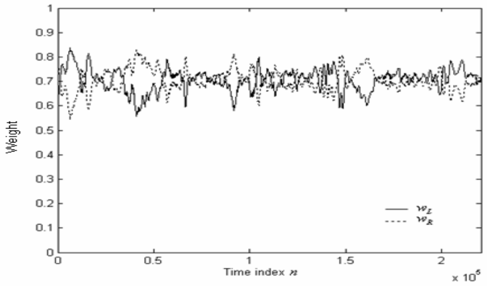


Fig. 5. The weights of the adaptive panning. The solid line is the panning weight corresponding to the left channel wL (n). The dashed line is the panning weight corresponding to the right channel wR (n).

D. The PCA-based Method

Along the line of the decorrelation idea, a PCA-based upmixing approach [4] is presented in this section. Let $w_L(n)$ and $w_R(n)$ be the left and the right channels. A 2×2 covariance matrix **A** is calculated as follows:

$$\mathbf{A} = \begin{bmatrix} cov(x_L, x_L) & cov(x_L, x_R) \\ cov(x_R, x_L) & cov(x_R, x_R) \end{bmatrix}, \tag{11}$$

where $cov(x_p,x_q)$, $p,q=L,R$ symbolizes the covariance estimated by an l-sample frame-based average. Specifically,

$$cov(x_L, x_R) = \sum_{n=1}^{l} [x_L(n) - \overline{x}_L][x_R(n) - \overline{x}_R] \big/ (l-1), \tag{12}$$

where $x_L(n)$ and $x_R(n)$ are the left and right signals at the time instant n, and $\overline{x}_L$ and $\overline{x}_R$ represent the means of the left and the right signals, respectively. The symmetric covariance matrix **A** guarantees to give two orthonormal eigenvectors. Let the eigenvectors associated with the larger and the smaller eigenvalues be (C_L,C_R) and (S_L,S_R), respectively. The center and the surround channels, presumably the correlated and the uncorrelated portions, can be derived by mixing the input signals according to the eigenvectors.

$$\text{Center} = C_L x_L(n) + C_R x_R(n) \tag{13}$$

$$\text{Surround} = S_L x_L(n) + S_R x_R(n) \tag{14}$$

Since the method is essentially frame-based processing, the discontinuities in mixing coefficients may lead to artifacts at the frame boundaries. A fading procedure is proposed to smooth the crossover of weights between successive frames. Let $C_L(k)$ and $C_L(k+1)$ be the coefficients to mix the left channel into the center channel in two successive frames, k and $k + 1$. Each frame is further divided into four smaller sub-frames. As shown in Fig. 6(a), the fading scheme proceeds with the following mixing coefficients, $w_{m,L\text{-}C}$, $m = 1,2,3,4$, for the sub-frames

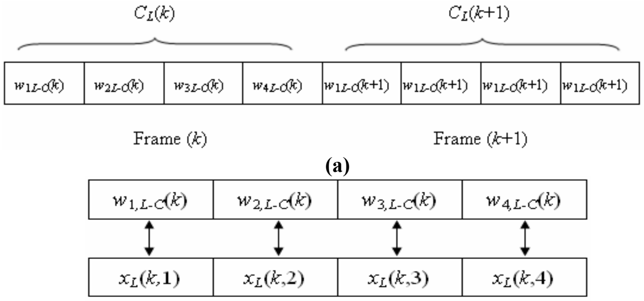
$$\begin{aligned} w_{1,L-C}(k) &= C_L(k) \\ w_{2,L-C}(k) &= 0.75 \times C_L(k) + 0.25 \times C_L(k+1) \\ w_{3,L-C}(k) &= 0.5 \times C_L(k) + 0.5 \times C_L(k+1) \\ w_{4,L-C}(k) &= 0.25 \times C_L(k) + 0.75 \times C_L(k+1) \end{aligned} \tag{15}$$

Figure 6(b) shows the correspondence between the mixing coefficients and the left channel signals. Finally, the center and surround channels are produced by mixing the front channels using the new mixing coefficients as follows:

$$\begin{aligned} C(k,m) &= w_{m,L-C} x_L(k,m) + w_{m,R-C} x_R(k,m) \\ S(k,m) &= w_{m,L-S} x_L(k,m) + w_{m,R-S} x_R(k,m) \\ m &= 1,2,3,4 \end{aligned} \tag{16}$$

E. The Reverb-based Method

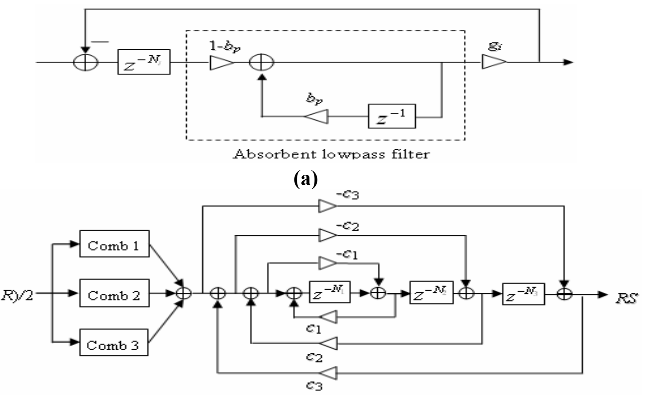
An artificial revereberator is employed in this section to produce the ambience-enriched surround channels [5]. The reverberator comprises three parallel comb filters (Fig. 7(a)) and three nested allpass filters (Fig. 7(b)) which are used to increase the modal density and echo density of reverberation. The parameters are determined by a sophisticated optimization procedure using the Genetic Algorithm (GA) [13]. In this approach, the surround channels are simply generated by feeding the average of the stereo inputs to the above-mentioned reverberator, plus 180-degree phase reversal.



(a)

(b)

Fig. 6. The interpolation procedure of the weights in the PCA-based upmixing technique. (a) Frame and sub-frame structure. (b) The correspondence of the weights and the left channel signal.



(a)

(b)

Fig. 7. The allpass-comb network of the reverberator. (a) The structure of each comb filter, where the *bp* is the gain of absorbent lowpass filter and *kp* is the gain of comb filter. (b) The reverb filter comprising 3 parallel comb filters and 3 serial nested allpass filters.

III. Downmixing Algorithms

In many applications such as personal computer (PC) multimedia systems, portable audio products, only two-channel stereo loudspeakers are available. Thus, downmixing is necessary to convert multichannel audio content into two channels for loudspeaker presentation. In what follows, two downmixing techniques will be presented.

A. The Standard Downmixing Method

The standard, ITU-R BS.775-1 [8], details how to downmix multichannel signals with simple gain adjustment. The architecture of such standard downmixing technique is shown in Fig. 8. That is,

$$\begin{aligned} L &= FL + 0.71 \times C + 0.71 \times RL \\ R &= FR + 0.71 \times C + 0.71 \times RR \end{aligned} \tag{17}$$

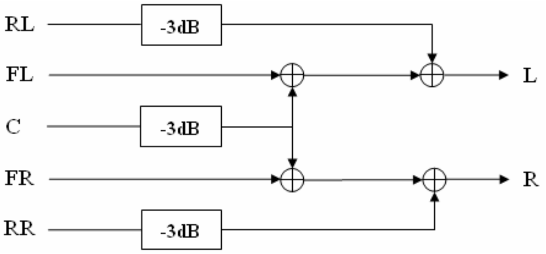


Fig. 8. The architecture of the standard downmixing method.

B. The HRTF-based Method

An HRTF-based downmixing technique is presented in the following. Figure 10 shows the architecture of the HRTF-based downmixing technique in which the rear surround channels are fed into a Shuffler filter [10]. For a symmetrical acoustical system, the Shuffler filter shown in Fig. 9 can be used to reduce computation cost. And Σ and Δ are given as

$$\Sigma = h_i + h_c \tag{18}$$

$$\Delta = h_i - h_c , \tag{19}$$

where h_i and h_c represent the ipsilateral and the contralateral HRTFs, respectively. In the HRTF-based method, the center, the rear left, and the rear right channels are filtered by the corresponding HRTFs at 0˚, +110˚, and −110˚, respectively, to provide directional impression before mixing with the front channels. The HRTF database implemented by using 128-tapped FIR filters is obtained from the website of the MIT media lab [7]. In the downmixing process, a crosstalk problem could arise as loudspeakers are used for reproducing the binaural signals. Excessive crosstalk could degrade sound localization especially for closely spaced loudspeakers. In order to alleviate the problem, crosstalk cancellation can be incorporated into the HRTF-based downmixing process. In this paper, a multichannel deconvolution method based on the Fast Fourier Transform (FFT) and Tikhonov regularization is adopted to calculate the frequency response functions of the CCS filters [10]:

$$\mathbf{C}(e^{j\omega}) = [\mathbf{H}^H(e^{j\omega})\mathbf{H}(e^{j\omega}) + \beta^2(\omega)\mathrm{I}]^{-1}\mathbf{H}^H(e^{j\omega}) \tag{20}$$

where $\mathbf{H}^H(e^{j\omega})$ is the hermitian transpose of the acoustical system $\mathbf{H}(e^{j\omega})$ and β is the regularization parameter. The coefficients of CCS filters can be obtained by applying inverse FFT and circular shifts to the frequency response functions.

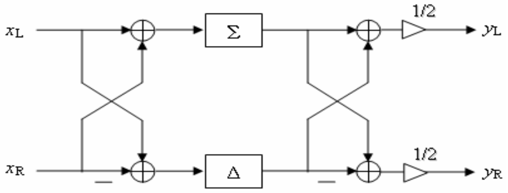


Fig. 9. The architecture of the Shuffler filter.

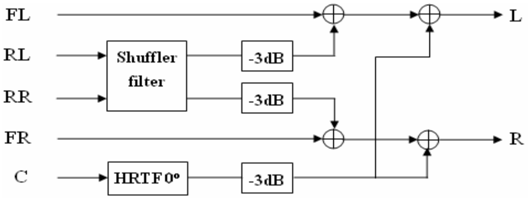


Fig. 10. The architecture of the HRTF-based downmixing method.

IV. Integration of UP/Downmixing algorithms

Most audio recordings now available are still in the stereo format and most consumer electronic products are equipped with two-channel loudspeakers. Upmixing and downmixing capabilities can be integrated to simulate a multichannel audio environment. The SRS 3D stereo sound system [6] is one such system with integrated upmixing and downmixing capabilities. In this paper, a hybrid method is introduced. The hybrid method processes the two-channel stereo inputs as follows:

$$Lout = K_0 L + K_1(L + R) + K_2(L - R)p \tag{21}$$

$$Rout = K_0 R + K_1(L + R) + K_2(R - L)p , \tag{22}$$

where (*L*+*R*) and (*L*-*R*) produce the upmixed center and the surround channels, K_0=1, K_1 =0.5, and K_2 =0.5 are the gains for the *L* and *R* channels, the center channel, and the surround channels, respectively. In the Shuffler filter, the HRTF filtering of the surround channels is only needed for the Δ filter since the sum of (*L*-*R*) and (*R*-*L*) equals zero. That is, the filter *p* in Eqs. (21) and (22) represents the difference of the ipsilateral and the contralateral transfer functions, (h_i- h_c). The above equations have combined the (Dolby-like) upmixing and dowmixing (with HRTFs) into one single step. An equalizer is employed to emphasize the lower frequencies (below 1 kHz) and the higher frequencies (7~20 kHz) and avoid the coloration problem resulting from the difference of signals.

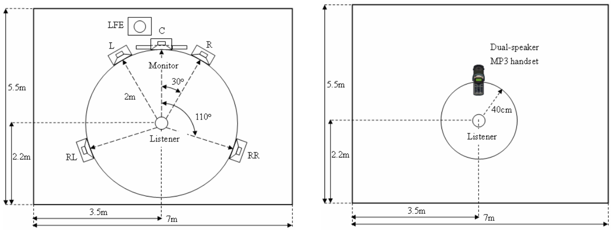
V. Subjective Evaluation of Upmixing and Downmixing Algorithms

A. Experimental Arrangement

1) Experimental Setup

Subjective listening tests were conducted to assess the performance of the aforementioned upmixing and downmixing algorithms. A 5.1 home theater system including five 3.5-inch loudspeakers and a subwoofer was used for sound reproduction.

The loudspeakers were deployed according to ITU-R BS.775 [8], as shown in Fig. 11(a). On the other hand, a dual loudspeakers MP3 handset was adopted in another subjective test, as shown in Fig. 11(b). The listening tests were conducted complying with the requirement of ITU-R BS.1116 [14]. As required in the CCS design, the binaural transfer functions from the loudspeakers to the microphone embedded in the ears of a KEMAR's were measured in an anechoic chamber by using a spectrum analyzer.



(a) (b)

Fig. 11. The experimental setup in the listening room. (a) The standard 5.1 configuration for multichannel loudspeaker reproduction. (b) The MP3 handset equipped with dual loudspeakers.

2) Procedure of Listening Test

A modified double-blind Multi-Stimulus test with Hidden Reference and a hidden Anchor (MUSHRA) (ITU-R BS.1534 [15]) was employed as the basis of the experimental design. The original unprocessed signal was used as the hidden reference in the test. The hidden anchor employed in this test was the 'phantom mono' reproduction that broadcasts the same signal over two-channel loudspeakers. The upmixing and downmixing algorithms were implemented on the platform of a fixed-point digital signal processor (DSP) equipped with a 6-input and 6-output codec operating at 48 kHz. The subjects are allowed to switch between different stimuli at their discretion and grade the presentations. The loudness of each reproduced signal was adjusted to equal level by a group of five skilled subjects to minimize experimental errors due to loudness variation. Three-leveled hierarchical subjective indices shown in Fig. 12 were employed to assess the performance of the upmixing and downmixing techniques.

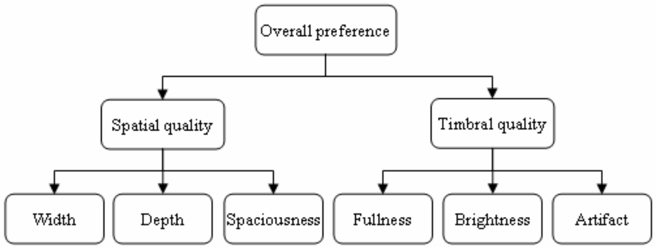


Fig. 12. Three-leveled hierarchical subjective indices employed in the listening test to assess the performance of the upmixing and downmixing techniques.

In the lowest hierarchy, the width, depth, and spaciousness refer to the perceived angular width, depth of the sound image, and the ambience, envelopment, and sensation of space pertaining to the listening environment. The fullness and brightness refer to the dominant low and high frequency content, respectively. The index, artifact, refers to any linear or nonlinear distortions. In the second hierarchy, the spatial quality and timbral quality describe the general performance of the spatial and timbral characteristics, respectively. In the top hierarchy, the overall preference represents the global impression of the reproduced program. The grading scale used for the subjective tests are shown in Tables I .

TABLE I
THE GRADING SCALE USED IN THE MUSHRA PROCESS.

Performance	Grade
Much better	3
Better	2
Slightly better	1
About the same	0
Slightly worse	-1
Worse	-2
Much worse	-3

3) Test Subjects

Thirty listeners who are experienced in audio evaluation took part in the listening tests. A training phase was arranged for the listeners prior to the formal test such that the listeners were thoroughly familiarized with the test facilities, the environment, and the grading process. Different items of lowpass processing, highpass processing, phantom mono, and multichannel 5.1 mode of presentation were demonstrated in the training phase.

4) Data Analysis

The results of the tests were processed by using the MANOVA. Both the mean and the 95% confidence intervals of the grades were shown in the analysis results. Cases with significance levels below p = 0.05 indicate that difference among methods is statistically significant. The scores obtained from the hidden reference and anchor were only intended for validating the consistency of subjects, but were excluded from the MANOVA as usually done in practice. Three statistical assumptions of MANOVA such as independence of grading, normal distribution of scores, and homogeneity of variances (HOV) were verified in these experiments. Independence of grading has been fulfilled due to the randomization of the experimental factors. The assumptions of normality and HOV were examined by using Shapiro-Wilk's W test and Levene's test, respectively [16]. Furthermore, a multiple regression model was employed to correlate the global auditory attributes with the low-level attributes in the listening test.

TABLE II
THE STATISTICAL ANALYSIS RESULTS OF THE LMS-BASED UPMIXING TEST FOR THE HOME THEATER LOUDSPEAKERS.

Dependent Variable	Type III SS	df	MS	*F*	*p*
Width	0.545455	2	0.272727	0.387931	0.681813
Depth	0.242424	2	0.121212	0.322581	0.726759
Spaciousness	0.424242	2	0.212121	0.472973	0.627715
Fullness	0.181818	2	0.090909	0.111940	0.894469
Brightness	0.000000	2	0.000000	0.000000	1.000000
Artifact	0.000000	2	0.000000	0.000000	1.000000
Spatial quality	0.424242	2	0.212121	0.813953	0.452649
Timbral quality	0.000000	2	0.000000	0.000000	1.000000
Overall preference	0.424242	2	0.212121	0.372340	0.692260

TABLE III
THE STATISTICAL ANALYSIS RESULTS OF THE UPMIXING TEST FOR THE HOME THEATER LOUDSPEAKERS.

Dependent Variable	Type III SS	df	MS	*F*	*p*
Width	20.72000	4	5.18000	2.749747	0.034795
Depth	1.41333	4	0.35333	0.216706	0.928302
Spaciousness	18.34667	4	4.58667	3.698925	0.008650
Fullness	7.06667	4	1.76667	1.069781	0.378068
Brightness	7.12000	4	1.78000	1.005920	0.410419
Artifact	13.78667	4	3.44667	5.337758	0.000825
Spatial quality	14.74667	4	3.68667	2.662311	0.039562
Timbral quality	21.46667	4	5.36667	4.551696	0.002517
Overall preference	41.94667	4	10.48667	4.942101	0.001442

B. *Evaluation of Upmixing Algorithms*

An experiment was conducted to compare the upmixing methods for PC home theater loudspeakers. Upmixing does not apply to the case of the handset because of the limited number of the rendering loudspeakers. We ran a listening test to choose the LMS-based approach that was most effective in upmixing. The result shown in Fig. 13 revealed that the correlation-based approach slightly outperformed the other methods in spaciousness and overall preference. However, the difference in performance was not significant because the *p*-values of the MANOVA output summarized in Table II were above 0.05. Therefore, we choose the simple LMS algorithm for upmixing because of its computational efficiency.

Figure 14 compares five upmixing methods by plotting the mean grades with 95% confidence intervals. The small *p*-values of MANOVA output were summarized in Table III. As expected, the grade of the hidden reference was nearly zero. The phantom mono attained the lowest grade except for depth, fullness, and artifact. This justifies the reliability and consistency of the test subjects. The reverb-based method outperforms the other methods in the spatial attributes, albeit ringing artifacts were reported by some subjects. The reverb-based method has attained the highest grade among all methods in overall preference. This seems to suggest that reverb-based upmixing methods are subjectively superior to the correlation-based methods.

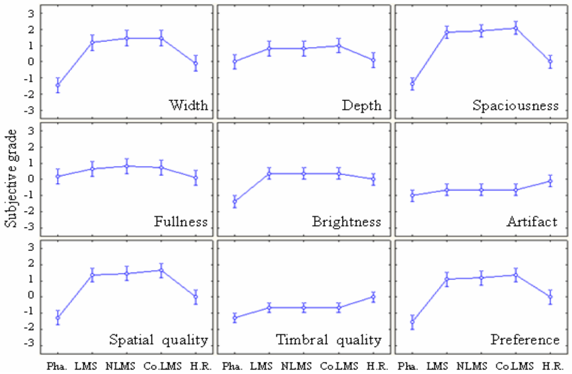


Fig. 13. Listening test results of three LMS-based upmixing methods for the home theater loudspeakers. (Pha: phantom mono reproduction, LMS: LMS algorithm, NLMS: NLMS algorithm, Co.LMS: correlation-based LMS method, H.R.: hidden reference).

C. *Evaluation of Downmixing Algorithms*

In this section, the downmix processing using the standard downmix method, the HRTF-based method, and the HRTF-CCS-based method will be examined. The 5.1 home theater system and a dual-loudspeaker MP3 handset are as rendering systems, respectively. In the case of home theater, the 5.1 setup was employed as the reference. In the case of handset, the standard downmix method was employed as the reference. The test results of the downmixing for home theater system are shown in Fig. 15. As expected, the grades of the hidden reference and phantom mono reproduction were quite low in most aspects.

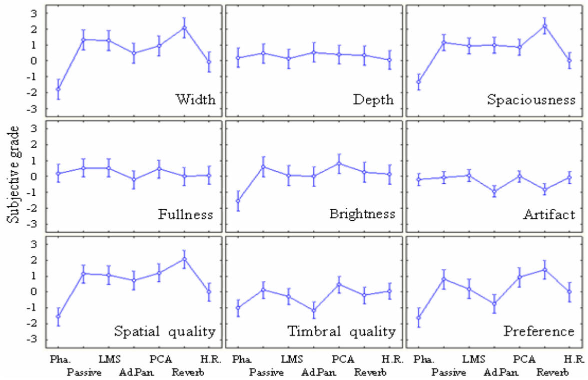


Fig. 14. Listening test results of five upmixing methods for the home theater loudspeakers. (Pha: phantom mono reproduction, Passive: passive surround decoder, LMS: LMS-based method, Ad.Pan.: adaptive panning method, PCA: PCA-based method, Reverb: reverb-based method, H.R.: hidden reference).

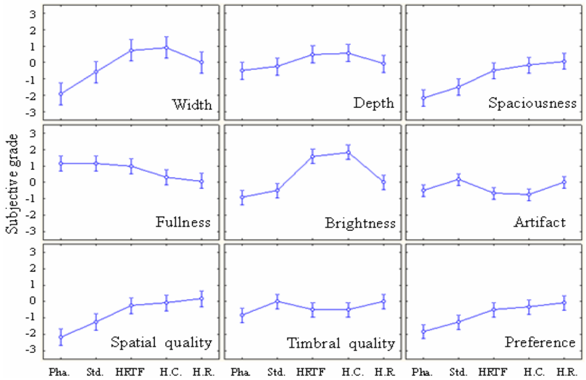


Fig. 15. Listening test results of three downmixing techniques for the home theater loudspeakers (Pha: phantom mono reproduction, Std.: standard downmixing, HRTF: HRTF-based method, H.C.: HRTF-CCS-based method, H.R.: hidden reference).

The MANOVA output shown in Table IV indicated significant difference among the approaches. The HRTF-CCS-based method has attained the higher grade in spatial characteristics and overall preference, but overlapping of the confidence intervals suggested that the difference was not statistically significant. The fact that the CCS did not work as expected, especially for widely spaced loudspeakers, could be due to several reasons [17] [18]. First, the reflections from boundaries may have obscured the localization of sound images. Second, CCS can increase only marginal performance for widely spaced loudspeakers because the natural separation is inherently good in such situation. Third, the CCS is

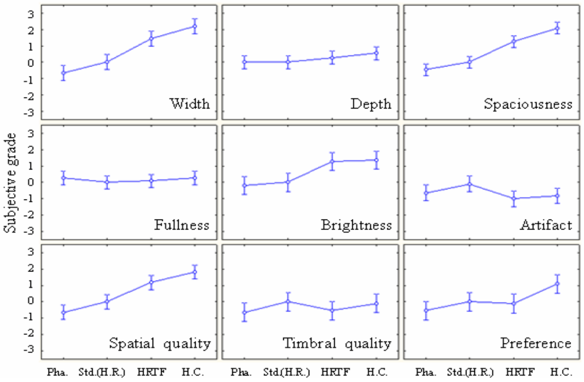


Fig. 16. Listening test results of three downmixing techniques for the MP3 handset, with the mean and 95% confidence intervals of scores indicated on the plot. (Pha: phantom mono reproduction, Std.(H.R.): standard downmixing referred as hidden reference, HRTF: HRTF-based method, H.C.: HRTF-CCS-based method).

TABLE IV
THE STATISTICAL ANALYSIS RESULTS OF THE FOWNMIXING TEST FOR THE HOME THEATER LOUDSPEAKERS.

Dependent Variable	Type III SS	df	MS	F	p
Width	16.22222	2	8.11111	4.17685	0.024154
Depth	5.05556	2	2.52778	2.59326	0.089942
Spaciousness	11.55556	2	5.77778	5.66337	0.007682
Fullness	4.66667	2	2.33333	2.53846	0.094310
Brightness	39.38889	2	19.69444	25.40391	0.000000
Artifact	6.16667	2	3.08333	6.97714	0.002971
Spatial quality	9.55556	2	4.77778	4.45176	0.019424
Timbral quality	2.00000	2	1.00000	1.65000	0.207501
Overall preference	5.72222	2	2.86111	4.30798	0.021762

TABLE V
THE STATISTICAL ANALYSIS RESULTS OF THE DOWNMIXING TEST FOR THE MP3 HANDSET.

Dependent Variable	Type III SS	df	MS	F	p
Width	27.15152	2	13.57576	32.94118	0.000000
Depth	1.63636	2	0.81818	1.64634	0.209694
Spaciousness	24.42424	2	12.21212	51.66667	0.000000
Fullness	0.42424	2	0.21212	0.48611	0.619774
Brightness	12.78788	2	6.39394	6.67722	0.003993
Artifact	5.09091	2	2.54545	4.11765	0.026294
Spatial quality	18.72727	2	9.36364	30.29412	0.000000
Timbral quality	1.87879	2	0.93939	1.01974	0.372853
Overall preference	9.51515	2	4.75758	5.13072	0.012119

ineffective as the listener moved outside the sweet spot for binaural reproduction, where the sweet spot of widely spaced loudspeakers is smaller than that of the closely spaced loudspeakers [18]. Figure 16 shows the test result of the downmixing for the handset. The MANOVA outputs were summarized in Table V. Among the methods, no significant difference was found in fullness because the handset loudspeakers literally had no sufficient low-frequency response. Although HRTF improved spatial quality, it also had detrimental effect on timbral quality. Consequently, the HRTF-based method was no longer superior to the standard downmixing method in overall preference as in the previous test. However, the CCS technique has significantly improved the spatial impression. This suggests that the use of CCS is critical to downmixing for closely-spaced loudspeakers.

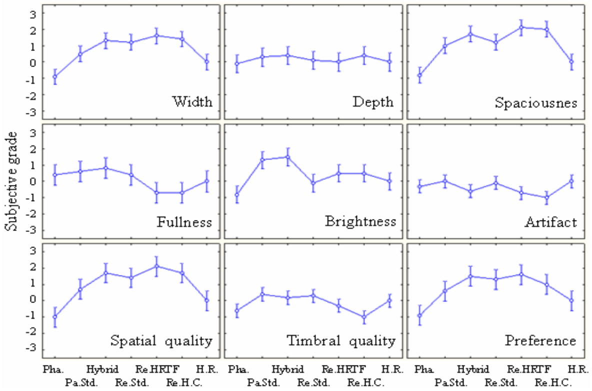


Fig. 17. Listening test results of up/downmixing techniques for the home theater loudspeakers. (Pha: phantom mono reproduction, Pa.Std.: passive surround decoder upmixing + standard downmixing, Hybrid: the hybrid method, Re.Std.: reverb-based upmixing + standard downmixing, Re.HRTF: reverb-based upmixing + HRTF-based downmixing, Re.H.C.: reverb-based upmixing + HRTF-CCS-based downmixing, H.R.: hidden reference).

D. Evaluation of Up/Downmixing Algorithms

Another listening test was conducted to evaluate the combined upmixing and downmixing techniques. In addition to the above-mentioned hybrid method, the procedure that integrates the passive surround decoder for upmixing and the standard method for downmixing was included in the test as a benchmark. Also, the reverb-based upmixing approach that has achieved the best performance in the preceding tests was integrated with various downmixing methods, including the standard downmixing method, the HRTF-based method, and the HRTF-CCS-based method. The front right and the front left loudspeakers of the home theater system were used as the rendering devices in the home theater test, whereas the two microspeakers in the handset were used as the rendering devices in the handset test.

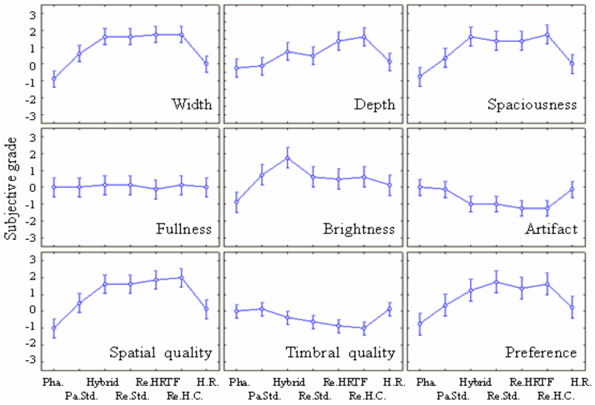


Fig. 18. Listening test results of the up/downmixing techniques for the MP3 handset. (Pha: phantom mono reproduction, Pa.Std.: passive surround decoder upmixing + standard downmixing, Hybrid: the hybrid method, Re.Std.: reverb-based upmixing + standard downmixing, Re.HRTF: reverb-based upmixing + HRTF-based downmixing, Re.H.C.: reverb-based upmixing + HRTF-CCS-based downmixing, H.R.: hidden reference).

TABLE VI
THE STATISTICAL ANALYSIS RESULTS OF THE UP/DOWNMIXING TEST FOR THE HOME THEATER LOUDSPEAKERS.

Dependent Variable	Type III SS	df	MS	*F*	*p*
Width	7.00000	4	1.750000	2.916667	0.031485
Depth	1.32000	4	0.330000	0.339041	0.850181
Spaciousness	9.40000	4	2.350000	3.975564	0.007591
Fullness	21.08000	4	5.270000	4.686759	0.003011
Brightness	17.12000	4	4.280000	4.755556	0.002757
Artifact	7.08000	4	1.770000	3.403846	0.016265
Spatial quality	10.88000	4	2.720000	2.467742	0.058232
Timbral quality	13.48000	4	3.370000	6.266529	0.000427
Overall preference	6.60000	4	1.650000	1.444553	0.234981

TABLE VII
THE STATISTICAL ANALYSIS RESULTS OF THE UP/DOWNMIXING TEST FOR THE MP3 HANDSET.

Dependent Variable	Type III SS	df	MS	*F*	*p*
Width	7.35000	4	1.837500	3.868421	0.010504
Depth	15.65000	4	3.912500	5.676166	0.001250
Spaciousness	9.40000	4	2.350000	3.576087	0.015115
Fullness	0.40000	4	0.100000	0.118644	0.974978
Brightness	8.35000	4	2.087500	2.731308	0.044484
Artifact	6.90000	4	1.725000	3.037736	0.029940
Spatial quality	11.35000	4	2.837500	4.815152	0.003356
Timbral quality	6.40000	4	1.600000	3.612903	0.014433
Overall preference	9.35000	4	2.337500	2.671429	0.048086

Figure 17 shows the test results of up/downmixing by using the home theater loudspeakers. The MANOVA output was summarized in Table VI. The benchmark method attained lower grade than the others in most attributes but timbral. The results showed that the hybrid method and the reverb-based upmixing combined with the HRTF-based downmixing method were preferred over the others. Surprisingly, however, the HRTF-CCS-based downmixing did not improve the performance over the HRTF-based downmixing, but decrease somewhat in both spatial and timbral qualities. Timbral quality degraded because CCS processing involved ipsilateral equalization which altered the perceived timbre. The gain in crosstalk cancellation seemed to be less than the loss in timbral quality when the CCS is applied to widely spaced loudspeakers. In any rate, the processed signals yielded predominantly better performance than the unprocessed reference, which justifies the necessity of up/downmixing in multichannel audio reproduction.

The result of up/downmixing for the MP3 handset is shown in Fig 18. The MANOVA output summarized in Table VII indicated a significant difference among the up/downmixing approaches. The results follow a similar trend to that of the home theater. The HRTF-based downmixing technique for the microspeakers did not perform as well as that in the home theater test due to the crosstalk problem. The improvement on spatial quality seemed to be totally offset by the degradation of timbral quality. However, addition of CCS to HRTF in downmixing seemed to have slightly recovered spatial quality. Similar to the home theater, the signals processed by the up/downmixing methods performed predominantly better than the unprocessed reference.

E. The regression analysis of auditory attibutes

The regression models are obtained as follows:

$$\text{Preference} = 0.545 \times \text{Spatial} + 0.568 \times \text{Timbral} + 0.048 \quad (R^2 = 0.587) \tag{23}$$

$$\text{Spatial} = 0.156 \times \text{Width} + 0.051 \times \text{Depth} + 0.726 \times \text{Spaciousness} + 0.077 \quad (R^2 = 0.785) \tag{24}$$

$$\text{Timbral} = 0.125 \times \text{Fullness} + 0.209 \times \text{Brightness} + 0.751 \times \text{Artifact} - 0.049 \quad (R^2 = 0.593) \tag{25}$$

The squared correlation coefficient (R^2) of the predicted results indicates the regression model is statistically significant. The coefficients in the model suggest the weighting that a low-level attribute contributes to the global attribute. The regression model revealed that the spatial and the timbral quality contributed comparably to the overall preference. In addition, spaciousness and artifact are dominant attributes to the spatial quality and the timbral quality, respectively. This suggests that the top and the second level of auditory attributes may have sufficed the listening test.

VI. CONCLUDING REMARKS

A comprehensive study has been carried out to compare various upmixing and downmixing techniques. A 5.1 home theater system and a dual-loudspeaker MP3 handset were used as the rendering devices. Subjective listening tests were conducted. Conclusions drawn from the results are summarized as follows.

Among the upmixing methods, the reverb-based technique has attained the best performance in spatial quality as well as in overall preference. This also suggests that the reverb-based method is subjectively more preferred in upmixing over the correlation-based methods.

For downmixing techniques applied to widely spaced loudspeakers, the HRTF-based method outperformed the standard downmixing method in the spatial quality and overall preference. The combined HRTF-CCS-based downmixing improved only marginally the performance, but the difference was not statistically significant. However, for closely spaced loudspeakers, the HRTF-based method was no longer superior to the standard downmixing method in overall preference as in the previous test. However, the CCS technique has significantly improved the spatial impression. This suggests that the use of CCS is critical to downmixing for closely-spaced loudspeakers.

For combined up/downmixing when applied to widely spaced loudspeakers, the hybrid method and the reverb-based upmixing combined with the HRTF-based downmixing method were preferred over the others. In particular, the latter approach has attained better spatial quality than the former approach. For methods using reverb-based upmixing, HRTF downmixing did have positive effects in spatial quality and overall preference. However, the HRTF-CCS-based downmixing did not improve the performance over the HRTF-based downmixing, but decrease somewhat in both spatial and timbral qualities. The gain in crosstalk cancellation seemed to be less than the loss in timbral quality when the CCS is applied to widely spaced loudspeakers. In combined use of up/downmixing for the microspeakers, downmixing using HRTF did not perform as well as that in the home theater test due to the crosstalk problem. The improvement on spatial quality seemed to be totally offset by the degradation of timbral quality. Under such circumstance, the CCS technique is crucial to minimize the crosstalk and recover spatial quality. In addition, the reverb-standard up/downmixing has achieved comparable preference with the reverb-HRTF-CCS up/downmixing because the former approach performed quite well in the timbral quality. Therefore, the up/downmixing method using reverberator and simple standard downmixing is preferred to realize virtual surround for handsets if computation cost is of concern. In any rate, the processed signals yielded predominantly better performance than the unprocessed reference, which justifies the necessity of up/downmixing in multichannel audio reproduction.

Acknowledgment

The work was supported by the National Science Council in Taiwan, Republic of China, under the project number NSC94-2212-E-009-019.

References

[1] Dolby Laboratory, Dolby surround Pro Logic decoder principles of operation, http://www.dolby.com/resources/tech_library/index.cfm

[2] B. Widrow and S.D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, 1985.

[3] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *J. Audio Eng. Soc.*, Vol. 50, pp. 914-926, 2002.

[4] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 2002.

[5] M. R. Bai and G. Bai, "Optimal design and synthesis of reverberators with a fuzzy user interface for spatial audio,“ *J. Audio Eng. Soc.*, Vol. 54, pp. 812-825, 2005.

[6] S. Gajjar, "A 3D stereo sound system," *IEEE Colloquium on Audio and Music Technology: The Challenge of Creative DSP, London*, pp. 15/1-15/7, 1998.

[7] B. Gardner and K. Martin, "HRTF measurements of KEMAR dummy-head microphone," MIT Media Lab, 1994, http://sound.media.mit.edu/KEMAR.html

[8] ITU-R BS.775-1, "Multi-channel stereophonic sound system with or without accompanying picture," International Telecommunications Union, Geneva, Switzerland, 1992–1994.

[9] W. G. Gardner, *3-D Audio Using Loudspeakers*. Kluwer Academic Publishers, 1998.

[10] O. Kirkeby, P. A. Nelson, and H. Hamada, "Fast deconvolution of multichannel systems using regularization," *IEEE Trans. Speech and Audio Processing*, Vol. 6, pp. 189-195, 1998.

[11] U. Zolzer, *DAFX, Digital Audio Effects*, John Wiley & Sons, 2002.

[12] P. Heitkamper, "An adaptation control for acoustic echo cancellers," *IEEE Signal Processing Letters*, Vol. 4, pp 170-172, 1997.

[13] R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*, 2nd ed., Wiley, 2004.

[14] ITU-R BS.1116, "Method of subjective assessment of small impairments in audio systems including multichannel sound systems," International Communications Union, Geneva, Switzerland, 1994.

[15] ITU-R BS.1534-1, "Method for the subjective assessment of intermediate sound quality (MUSHRA)", International Telecommunications Union, Geneva, Switzerland, 2001.

[16] D. C. Howell, *Statistical Method for Psychology*, Duxbury, NY, 1997.

[17] M. R. Bai, C. W. Tung, and C. C. Lee, "Optimal design of loudspeaker arrays for robust cross-talk cancellation using the Taguchi method and the genetic algorithm," *J. Acoust. Soc. Am.*, Vol. 117, pp. 2802-2813, 2005.

[18] M. R. Bai and C. C. Lee, "Objective and subjective analysis of effects of listening span on crosstalk cancellation in spatial sound reproduction," *J. Acoust. Soc. Am.*, to appear in Sept. 2006.

Mingsian R. Bai was born in 1959 in Taipei, Taiwan, ROC. He received a bachelor's degree in Power Mechanical Engineering from National Tsing-Hwa University in 1981. He also received a master degree in Business Management from National Chen-Chi University in 1984. He left Taiwan in 1984 to enter graduate school of Iowa State University and later received a MS degree from Mechanical Engineering in 1985 and a Ph. D. from Engineering Mechanics and Aerospace Engineering in 1989. In 1989, he joined the Department of Mechanical Engineering of National Chiao-Tung University in Taiwan as an associate professor and became a professor in 1996. He was also a visiting scholar to Center of Vibration and Acoustics, Penn State University, University of Adelaide, Australia, and Institute of Sound and Vibration Research (ISVR), UK in 1997, 2000, 2002, respectively. His current interests encompass acoustics, audio signal processing, electroacoustic transducers, vibroacoustic diagnostics, active noise and vibration control, and so forth. He currently serves as an active consultant and project leader in these areas in industry. He has over 100 published papers and 13 granted or pending patents. Professor Bai is a member of the *Audio Engineering Society (AES), Acoustical Society of America (ASA), Acoustical Society of Taiwan, and Vibration and Noise Control Engineering Society in Taiwan.*

Geng-Yu Shih was born in 1982 in Tainan, Taiwan, ROC. He received a bachelor's degree in Mechanical Engineering from the National Central University in 2004. He is currently working on the Mechanical Engineering master degree in the National Chiao-Tung University. His master thesis is on implementation of a 3D audio module with up/downmix and CCS for two-channel stereo loudspeakers.