Enhancement of Spatial Sound Quality: A New Reverberation-Extraction Audio Upmixer

John Usher, *Member, IEEE*, and Jacob Benesty, *Senior Member, IEEE*

***Abstract*—A system for the extraction of uncorrelated reverberation from two-channel (stereo) audio signals is proposed and evaluated. Applications for the new system vary from surround-sound multichannel loudspeaker upmixers for home-theater or automotive audio systems, to headphone-based auralization for enhancing the spatial sound quality of a listening experience. The new system uses the normalized-least-mean-square (NLMS) algorithm to equalize the two input signals with respect to both spectral magnitude and phase before differencing to remove correlated components. A theoretical model of the system based on a stochastic room impulse response model was validated by empirical measurements made in a reverberant hall with a microphone pair, and from a formal subjective evaluation the system is shown to be an effective approach to extracting reverberation from audio recordings.**

***Index Terms*—Adaptive filtering, ambiance extraction, loudspeakers, reverberation, sound quality, spatial audio, upmixing.**

I. Introduction

EXTRACTION of reverberation from audio signals is an increasingly relevant problem for spatial audio system designers. Upmixing two-channel ("stereo") audio recordings to four or five channels allows for reproduction with immersive multichannel loudspeaker systems found in domestic "home theater," automotive audio, and teleconferencing environments. Furthermore, headphone audio systems would benefit from spatial audio enhancement with ambiance auralization processors to reproduce the recorded reverberation in a way that seems more enveloping. These "ambiance extraction" upmixers can reduce the number of channels stored or transmitted, while increasing the spatial sound quality of the recreated sound scene. Recent audio upmixers [1]–[4] rely on a common principle of extracting reverberation embedded within the recording, rather than adding artificial decorrelation. The method for accomplishing this relies on the assumption that those sound components that affect our perception of recorded reverberation (i.e., reverberance) are uncorrelated in the two input channels, and that removal of the correlated sound components will yield the recorded reverberation. Removal of correlated sound components from the input signals is generally undertaken using a bootstrapped adaptive gain control mechanism to equalize the input signals in terms of level before differencing. As with the new upmixer presented here, two of these systems [3], [4] use a least-mean-square (LMS)-based algorithm to determine the gain of a principle sound source in each input channel on a frequency-by-frequency basis. These two systems first separate the input signals into different frequency bands and then apply a single gain coefficient to each channel before a difference signal is calculated. This approach can be problematic when the input signals are maximally correlated at a lag $\neq 0$, i.e., when the direct sound from the recorded sound source arrives in one channel before the other; common in sound recordings made with multiple spaced microphones (which applies to nearly all music-audio). For such situations, the extant ambiance extraction upmixers cannot effectively cancel the correlated sound components, and leakage of direct-sound components to the ambiance-channel occurs. The proposed new system in the present work overcomes this limitation using a frequency-domain implementation of a normalized-least-mean-square (NLMS) adaptive filter to align the input signals both spectrally and temporally before differencing, thereby allowing the correlated components to be removed. A formal listening test evaluating if the new upmix system provides a subjective improvement over conventional two-loudspeaker audio scenes is also reported.

Manuscript received November 5, 2006; revised May 14, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rudolf Rabenstein.

J. Usher was with McGill University, Montréal, QC H3A 1E3, Canada. He is now with Personics, Montreal, QC H2Y 1V7, Canada (e-mail: jusher@personicslabs.com).

J. Benesty is with Université du Québec, INRS-EMT, Montréal, QC H5A 1K6, Canada (e-mail: benesty@emt.inrs.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.901832

II. New Reverberation Extraction System

In the description of the new upmix system in Fig. 1, we will assume that the two input signals are directly from a microphone pair; therefore, the recording media can be eliminated from the discussion. These two signals at sample time n are $m_i(n)$ and $m_j(n)$ (where i or j =1 or 2 and $i \neq j$). For the generation of each output channel the upmix process can be summarized as follows:

1) filtering the m_i input audio signal with respect to a set of filtering coefficients (typically with a 1024-tap finite-impulse response (FIR) filter);[1]
2) time-shifting the m_j input audio signal with respect to the other input signal (typically with a delay of about 10 ms);
3) calculating a difference between the filtered m_i and time-shifted m_j signal. (This difference signal (y_j) can then be radiated with a separate loudspeaker, such as a rear loudspeaker in a multichannel audio system, or processed using

[1]In all experiments reported here, the sample rate of the digital signal processing system was 44.1 kHz.

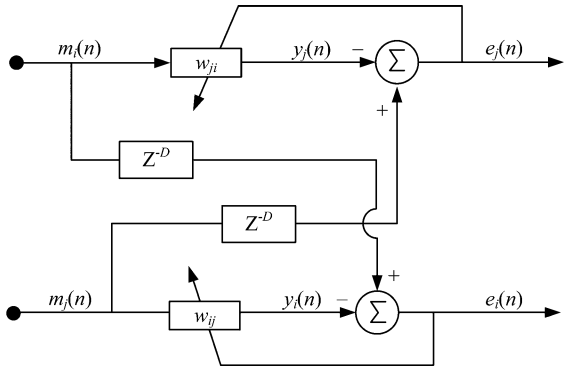

Fig. 1. Overview of the proposed ambiance extractor upmix system. The pair of input signals (e.g., from two microphones) at sample time n is $m_i(n)$ and $m_j(n)$ and are filtered with the adaptive filters w_{ji} and w_{ij}. The filters are adapted over time so that the level of the difference (error) signals are minimized, and these signals are radiated, for example, by additional loudspeakers behind the listener. The pure delay D is to allow for nonminimum phase impulse responses to allow the filtered input signal to be time delayed or advanced relative to the nonfiltered signal.

head-related transfer functions to increase the spatial fidelity of a headphone audio system);

4) adjusting the set of filtering coefficients so that the difference signal is essentially orthogonal to the m_i input signal.

A delay D is added to input channel m_i before the filtered signal y_i is subtracted; this is to allow for time-alignment of the two input signals if the direct sound arrives in one channel before the other. For the spectral and temporal equalization of nonminimum-phase impulse responses in audio systems, the Simpson Sideways Recursion technique has been used [5] within an LMS algorithm. This has the effect of introducing a constant delay at all frequencies to the equalizing filter response. While this approach may be effective for a single sound source, upmixing audio containing multiple recorded sound sources may require different temporal equalization at different frequencies. The delay D was approximately 10 ms because this allows for frequency-dependent temporal equalization up to the maximum interchannel time delay possible with microphones spaced by up to 3.4 m; which is a larger spacing than is generally used in music recordings [6].

The MINT principle for a single source and multiple microphones [7] has a different goal than our approach. Indeed, in this paper we try to extract reverberation and use it for a better sound reproduction with multiple loudspeakers while the MINT method attempts to remove reverberation of the room.

III. Notation and Adaptive Algorithms

A. Definitions and Notation

As discussed, in order to extract only those sound components that affect reverberance imagery, the ambiance extractor must remove those components in the two signals which are correlated. Therefore, the purpose of the adaptive filter can be interpreted as finding a function which can be applied to one input signal to make it electronically the same as the other (or as similar as possible), allowing the correlated sound components to be removed by subtracting the filtered signal from the other input signal.

The filtering of signal m_j by the adaptive filter $\mathbf{w}_{ij}$ gives signal $y_i(n)$. This filtering process is summarized as the discrete time linear convolution as follows:

$$y_i(n) = \sum_{k=0}^{M-1} m_j(n-k) w_{ij,k} \tag{1}$$

which with the following definitions:

- $\mathbf{m}_j(n) = [m_j(n), m_j(n-1), \ldots, m_j(n-M+1)]^T$.
- $\mathbf{w}_{ij} = [w_{ij,0}, w_{ij,1}, \ldots, w_{ij,M-1}]^T$.

allow the convolution to be written in vector form as

$$y_i(n) = \mathbf{m}_j^T(n)\mathbf{w}_{ij}. \tag{2}$$

Looking at filter system $\mathbf{w}_{ij}$ in Fig. 1, the filtered m_j signal y_i is subtracted from the unfiltered m_i signal (sample-by-sample) to give the error signal e_i

$$e_i(n) = m_i(n-D) - y_i(n). \tag{3}$$

To simplify the notation, we will consider that the delay $D = 0$, but in all experiments reported herein the delay is 500 samples (11 ms).

B. Optimization Criterion

If the filter coefficients $\mathbf{w}_{ij}$ can be adapted so as to approximate the intermicrophone impulse response, then the correlated sound component will be removed and the "left-over" signal e_j will contain the reverberant (or reverberance-image) component in the m_j channel. In this case, the error signal level e_j will be smaller than the original level of m_j. The "goal" of the algorithm that changes the adaptive filter coefficients can therefore be interpreted as to minimize the level of the error signals. This goal is formally expressed as a "performance index" or "cost" scaler J, where for a given filter vector $\mathbf{w}_{ij}$

$$J_i(\mathbf{w}_{ij}) = E\left\{e_i^2(n)\right\} \tag{4}$$

where $E\{\cdot\}$ is the statistical expectation operator.

The requirement for the algorithm is to determine the operating conditions for which J attains its minimum value: this state of the adaptive filter is called the "optimal state" [8].

C. Adaptation Algorithm

Updating the filter vector $\mathbf{w}_{ij}(n)$ from time $n-1$ to time n is done by multiplying the negative of the gradient operator by a constant scaler μ

$$\mathbf{w}_{ij}(n) = \mathbf{w}_{ij}(n-1) + \mu \mathbf{m}_j(n) e_i(n). \tag{5}$$

From the filter update (5), it can be seen that the adjustment from $\mathbf{w}_{ij}(n-1)$ to $\mathbf{w}_{ij}(n)$ is proportional to the filtered input vector $\mathbf{m}_j(n)$. When the filter has converged to the optimal solution, the rate of change of filter coefficients ($\boldsymbol{\nabla}$) should be zero but the actual $\boldsymbol{\nabla}$ will be equal to $\mu \mathbf{m}_j(n) e_i(n)$. This product may be not equal to zero and results in gradient noise [9] which

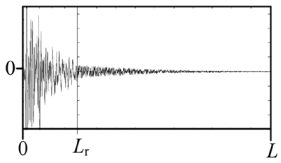

Fig. 2. Two-component acoustic impulse response model for the time-domain transfer function between two locations in a room. The x-axis is time and the y axis is pressure (or voltage). The first part up to the *mixing time* L_r is called the early component and consists of the direct sound and early reflections, which primarily affects the perceptual image corresponding to the recorded sound source (the source image). The second part is called the reverberant component, which primarily affects reverberance imagery.

is proportional to the level of $\mathbf{m}_j(n)$. This undesirable consequence can be mitigated by normalizing the gradient estimation with another scaler that is inversely proportional to the power of $\mathbf{m}_j(n)$, and the algorithm is therefore called the NLMS algorithm [10]

$$\mathbf{w}_{ij}(n) = \mathbf{w}_{ij}(n-1) + \frac{\alpha}{\delta + \mathbf{m}_j^T(n)\mathbf{m}_j(n)}\mathbf{m}_j(n)e_i(n) \quad (6)$$

with $0 < \alpha < 1$, and δ is a regularization constant added to the power estimate to ensure against computational problems for low input levels.

Besides the decrease in computational load of implementing the filter-update and signal filtering in the frequency domain [requiring five fast Fourier transforms (FFTs) per iteration], the performance of the frequency-domain and time-domain NLMS algorithm are equivalent [11]. In the present work, the overlap-save technique [12] was used with an overlap factor of 4.

D. Principle of Orthogonality

The optimal state is attained when the gradient operator is equal to zero, so under these conditions

$$E\{\mathbf{m}_j(n)e_i(n)\} = \mathbf{0}_{M\times 1}. \quad (7)$$

This last statement represents the principle of orthogonality [8, p. 96]. This is particularly advantageous if the system is to be used as an upmixer for multichannel loudspeaker systems with a pair of rear loudspeakers radiating the error signals, where the fidelity of the virtual image corresponding to the location of the recorded sound source (the *source image*) should be maintained by the upmixing process [13]. This is because the source image is only affected by correlated components within m_i and m_j[14], so a radiated signal which is uncorrelated with either m_i or m_j cannot contain a sound component that affects this source image.

IV. Signal Model

A. Room Impulse Response Model

The time-domain acoustic transfer function between two locations in an enclosed space—such as between a sound source and a microphone diaphragm—can be modeled as a two-part impulse response (IR) [15], as summarized in Fig. 2.

In this IR model, the L-length acoustic IR is represented as two decaying time sequences; one of which is defined between sample times $n = 0$ and $n = L_r - 1$, the other between $n = L_r$ and $n = L$. The first of these sequences represents the IR from the direct sound and early-reflections (ERs), and the other sequence represents the reverberation (the latter component is therefore called the reverberant component). ERs are defined as "those reflections which arrive ... via a predictable, non-stochastic directional path," [16] whereas reverberation is generally considered to be sound reflections impinging on a point (e.g., microphone) which can be modeled as an exponentially decaying, ergodic, stochastic process, with a Gaussian distribution and a mean of zero [17], [18]. Empirical validation of this IR model from concert hall measurements is reported in [13].

It is the early component of the IR that primarily affects source imagery, such as perceived source direction, width and distance, and the reverberant component that affects reverberance imagery, such as envelopment and feeling for the size of the room [19]. The reverberant component is created by a high temporal density of independently distributed discrete reflections. According to the central limit theorem, the local pressure distribution in a reverberant field is therefore normal (Gaussian) [17]. This allows the time boundary between these two components in our IR model (called the *mixing time*) to be empirically determined using a measure for normality such as kurtosis. This can therefore be used to identify the optimum filter length for removing correlated sound components, as shall be discussed in Section VI-B.

The input signals $m_i(n)$ and $m_j(n)$ are described by the acoustic convolution between the sound source signal $s(n)$ and the L_r-length direct-path coefficients summed with the convolution of $s(n)$ with the $(L - L_r)$-length reverberant-path coefficients, as shown as follows:

$$m_i(n) = \sum_{k=0}^{L_r-1} s(n-k)d_{i,k} + \sum_{l=L_r}^{L} s(n-l)r_{i,l}, \qquad i = 1 \text{ or } 2. \quad (8)$$

As mentioned, the direct-path IR coefficients are the first L_r samples of the L-length IR between the source and two microphones, and the reverberant-path IR coefficients are the remaining $(L-L_r)$ samples of these IRs. The time-varying source samples and time-invariant IRs are now defined as the vectors

- $\mathbf{s}_d(n) = [s(n), s(n-1), \ldots, s(n-L_r+1)]^T$.
- $\mathbf{s}_r(n) = [s(n-L_r), s(n-L_r-1), \ldots, s(n-L)]^T$.
- $\mathbf{d}_i = [d_{i,0}, d_{i,1}, \ldots, d_{i,L_r-1}]^T$.
- $\mathbf{r}_i = [r_{i,0}, r_{i,1}, \ldots, r_{i,L-L_r-1}]^T$.

$\mathbf{s}_r(n)$ is that part of the source signal that travels along the reverberant paths. The acoustic convolutions between the radiated acoustic source and the early and reverberant-path IRs in (8) can be written as

$$m_i(n) = \mathbf{s}_d^T(n)\mathbf{d}_i + \mathbf{s}_r^T(n)\mathbf{r}_i. \quad (9)$$

For convenience, the early and reverberant path convolutions are replaced with

$$s_{d,i}(n) = \mathbf{s}_d^T(n)\mathbf{d}_i$$

and

$$s_{r,i}(n) = \mathbf{s}_r^T(n)\mathbf{r}_i. \quad (10)$$

Thus, (9) becomes

$$m_i(n) = s_{d,i}(n) + s_{r,i}(n). \tag{11}$$

B. Assumptions

With the following definitions for the last L samples of the early and reverberant path sound arriving at time n:

- $\mathbf{s}_{d,i}(n) = [s_{d,i}(n), s_{du,i}(n-1), \ldots, s_{d,i}(n-L+1)]^T$.
- $\mathbf{s}_{r,i}(n) = [s_{r,i}(n), s_{r,i}(n-1), \ldots, s_{r,i}(n-L+1)]^T$.

The following assumptions about these early and reverberant path sounds are expressed using the statistical expectation operator $E\{\cdot\}$.

- The early part of both IRs ("direct-path") are at least partially correlated as follows:
  $E\left\{\mathbf{d}_i^T(n)\mathbf{d}_j(n)\right\} \neq 0;$
  $E\left\{\mathbf{s}_{d,i}^T(n)\mathbf{s}_{d,j}(n)\right\} \neq 0.$
- The late part of each IR (the "reverberant path") are uncorrelated with each other as follows:
  $E\left\{\mathbf{r}_i^T(n)\mathbf{r}_j(n)\right\} = 0;$
  $E\left\{\mathbf{s}_{r,i}^T(n)\mathbf{s}_{r,j}(n)\right\} = 0.$
- The two reverberant path IRs are uncorrelated with both early parts as follows:
  $E\left\{\mathbf{r}_i^T(n)\mathbf{d}_i(n)\right\} = 0;$
  $E\left\{\mathbf{s}_{r,i}^T(n)\mathbf{s}_{d,i}(n)\right\} = 0.$
- The reverberant path IR is decaying random noise with a normal distribution and a mean of zero as follows:
  $E\{\mathbf{r}_i(n)\} = 0;$
  $E\{\mathbf{s}_{r,i}(n)\} = 0.$

C. Validity of Assumptions

Thus far, the effect of room modes or resonances has been ignored. These occur due to reflections normal to the room boundaries which constructively interfere with itself. At low frequencies, the distance between the maxima of the modes will be large (i.e., equal to the wavelength). The frequency $f_{\text{Schroeder}}$ (in hertz) where we can assume a high modal overlap, and therefore a stochastic model of reverberation, is related to the -60 dB reverberation time RT_{60} (in seconds) and can be approximated as follows [20]:

$$f_{\text{Schroeder}} \approx 2000\sqrt{\frac{RT_{60}}{V}}\ \text{(Hz)} \tag{12}$$

where V is the room volume in m^3.

The stochastic model for the reverberant component of the IR is therefore only valid after the mixing time and for frequencies above the Schroeder frequency; as summarized in Fig. 3.

The pressure correlation between two random locations in a concert hall is unpredictable below the Schroeder frequency: the measurement locations may both be located on a modal maximum (increasing the correlation), or on a maxima and minima (giving a negative correlation), or may be uncorrelated due to modal interference. Above the Schroeder frequency, however, the spatial correlation function in a reverberant (diffuse) soundfield can be predicted as follows [21]:

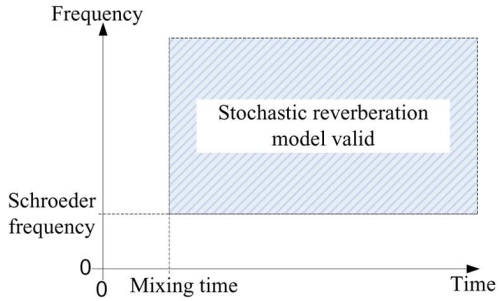

Fig. 3. Validity domain of the stochastic IR model in the time-frequency plane (hatched) [15]. The time axis corresponds to the IR sample time.

$$\kappa = \frac{\sin(kx)}{kx} \tag{13}$$

where the correlation between the two locations (κ) is dependant on the wavenumber[2] k and the distance between them x. Therefore, the assumption that the reverberant-path components are uncorrelated, i.e., $E\left\{\mathbf{s}_{r,i}^T(n)\mathbf{s}_{d,i}(n)\right\} = 0$ for a typical recording with a 20-cm spaced microphone pair is only valid for frequencies greater than approximately 1 kHz.

V. System Model

A. Effect of Input Correlation on Output Level

We define the normalized mean-squared error (NMSE) with respect to the reference signal as

$$\Psi_i = \frac{E\left\{e_i^2(n)\right\}}{E\left\{m_i^2(n)\right\}}. \tag{14}$$

Given a block of the last M samples of channel m_j ($\mathbf{m}_j(n)$) and the error (i.e., difference) signal

$$e_i(n) = m_i(n) - \mathbf{m}_j^T(n)\mathbf{w}_{ij} \tag{15}$$

the Wiener–Hopf equations [8] describe the relationship of the filter vector $\mathbf{w}_{ij}$ to the interinput channel correlation and the autocorrelation of channel m_j for a time-invariant optimum filter solution (i.e., when the error signal energy is minimized) and stationary source signals

$$\mathbf{w}_{ij} = \mathbf{R}_{m_jm_j}^{-1}\mathbf{r}_{m_jm_i} \tag{16}$$

where

$$\mathbf{r}_{m_jm_i} = E\{\mathbf{m}_j(n)m_i(n)\} \tag{17}$$

is the M-length cross-correlation vector, and

$$\mathbf{R}_{m_jm_j} = E\left\{\mathbf{m}_j(n)\mathbf{m}_j^T(n)\right\} \tag{18}$$

is the M-by-M autocorrelation matrix for signal m_j.

[2]For a frequency f, the wavenumber $k = 2\pi f/c$, where c is the speed of sound.

Using the definitions for the cross-correlation vector $\mathbf{r}_{m_jm_i}$ and auto-correlation matrix $\mathbf{R}_{m_jm_j}$, $E\left\{e_i^2(n)\right\}$ can be expressed as

$$E\left\{e_i^2(n)\right\} = \sigma_{m_i}^2 - 2\mathbf{r}_{m_jm_i}^T\mathbf{w}_{ij} + \mathbf{w}_{ij}^T\mathbf{R}_{m_jm_j}\mathbf{w}_{ij} \quad (19)$$

where $\sigma_{m_i}^2$ is the variance of the signal $m_i(n)$

$$\sigma_{m_i}^2 = E\left\{m_i^2(n)\right\}. \quad (20)$$

Now, substituting the definition of $\mathbf{w}_{ij}$ given in (16), we obtain from (19)

$$E\left\{e_i^2(n)\right\} = \sigma_{m_i}^2 - \mathbf{r}_{m_jm_i}^T\mathbf{R}_{m_jm_j}^{-1}\mathbf{r}_{m_jm_i}. \quad (21)$$

This allows Ψ_i, defined in (14), to be obtained by dividing (21) by $\sigma_{m_i}^2$

$$\Psi_i = 1 - \frac{\mathbf{r}_{m_jm_i}^T\mathbf{R}_{m_jm_j}^{-1}\mathbf{r}_{m_jm_i}}{\sigma_{m_i}^2} \quad (22)$$

which can be simplified by defining the normalized cross-correlation coefficient between $\mathbf{m}_j$ and m_i as the vector $\mathbf{c}_{m_jm_i}$[22]

$$\mathbf{c}_{m_jm_i} = \frac{\mathbf{r}_{m_jm_i}}{\sigma_{m_i}\sigma_{m_j}} \quad (23)$$

and (22) can be expressed as

$$\Psi_i = 1 - \sigma_{m_j}^2\mathbf{c}_{m_jm_i}^T\mathbf{R}_{m_jm_j}^{-1}\mathbf{c}_{m_jm_i}. \quad (24)$$

When $m_j(n)$ is white noise, the autocorrelation matrix $\mathbf{R}_{m_jm_j}$ is diagonal such that

$$\mathbf{R}_{m_jm_j} = \sigma_{m_j}^2\mathbf{I} \quad (25)$$

where $\mathbf{I}$ is the $M \times M$ identity matrix. So for near-white input signals, Ψ_i approximates

$$\begin{aligned}\Psi_i &= 1 - \frac{\sigma_{m_i}^2\sigma_{m_j}^2\mathbf{c}_{m_jm_i}^T\mathbf{c}_{m_jm_i}}{\sigma_{m_i}^2\sigma_{m_j}^2} \\ &= 1 - \mathbf{c}_{m_jm_i}^T\mathbf{c}_{m_jm_i}. \end{aligned} \quad (26)$$

Allowing Ψ_i to be conveniently expressed as a function of the two-norm of the cross-correlation vector $\mathbf{c}_{m_jm_i}$

$$\Psi_i = 1 - \left|\left|\mathbf{c}_{m_jm_i}\right|\right|^2. \quad (27)$$

B. Effect of Microphone Location on Output Level

The direct-to-reverberant (or reverberant-to-direct) ratio is a commonly used term in acoustics defined as the relative energy densities of the early-arriving and reverberant sound [23, p. 342]. The distance from the source at which this ratio is equal to unity is called the reverberation distance V (or radius) and for a room of volume (m^3) with reverberation time T_{60} (seconds), the distance r_h (meters) is approximated by (28) [24]

$$r_h = 0.1\sqrt{\frac{V}{\pi T}}. \quad (28)$$

All sound arriving after the direct sound up to the mixing time is included in the calculation of the direct sound energy and the reverberant-to-direct energy ratio γ_i is defined according to (29)

$$\gamma_i = \frac{E\left\{s_{r,i}^2(n)\right\}}{E\left\{s_{d,i}^2(n)\right\}}. \quad (29)$$

Conveniently, Ψ_i can be defined in terms of just γ_i and the direct-path correlation $\mathbf{c}_{s_{d,i}s_{d,j}}$ with a five-step procedure.

1) Starting with (22) and expanding the cross-correlation vector and autocorrelation matrix gives

$$\begin{aligned}\Psi_i = 1 &- E\left\{\mathbf{m}_j(n)m_i(n)\right\}^T \\ &\times \left(E\left\{m_i^2(n)\right\}E\left\{\mathbf{m}_j(n)\mathbf{m}_j^T(n)\right\}\right)^{-1} \\ &\times E\left\{\mathbf{m}_j(n)m_i(n)\right\}. \end{aligned} \quad (30)$$

2) Using the definitions for the signals $\mathbf{m}_j(n)$ and $m_i(n)$ allows (30) to be written as

$$\begin{aligned}\Psi_i = 1 &- E\left\{\left(\mathbf{s}_{d,j}(n) + \mathbf{s}_{r,j}(n)\right)\left(s_{d,i}(n) + s_{r,i}(n)\right)\right\}^T \\ &\times \left(E\left\{\left(s_{d,i}(n) + s_{r,i}(n)\right)^2\right\}\right. \\ &\qquad \left.\times E\left\{\left(\mathbf{s}_{d,j}(n) + \mathbf{s}_{r,j}(n)\right)\left(\mathbf{s}_{d,j}^T(n) + \mathbf{s}_{r,j}^T(n)\right)\right\}\right)^{-1} \\ &\times E\left\{\left(\mathbf{s}_{d,j}(n) + \mathbf{s}_{r,j}(n)\right)\left(s_{d,i}(n) + s_{r,i}(n)\right)\right\}. \end{aligned} \quad (31)$$

3) Expanding (31), we get

$$\begin{aligned}\Psi_i = 1 &- \left(E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\} + \underline{E\left\{\mathbf{s}_{d,j}(n)s_{r,i}(n)\right\}}\right. \\ &\qquad \left. + \underline{E\left\{\mathbf{s}_{r,j}(n)s_{d,i}(n)\right\}} + \underline{E\left\{\mathbf{s}_{r,j}(n)s_{r,i}(n)\right\}}\right)^T \\ &\times \left(\left(E\left\{s_{d,i}^2(n)\right\} + \underline{2E\left\{s_{d,i}(n)s_{r,i}(n)\right\}}\right.\right. \\ &\qquad \left.\left. + E\left\{s_{r,i}^2(n)\right\}\right) \\ &\qquad \times \left(E\left\{\mathbf{s}_{d,j}(n)\mathbf{s}_{d,j}^T(n)\right\} + \underline{E\left\{\mathbf{s}_{d,j}(n)\mathbf{s}_{r,j}^T(n)\right\}}\right. \\ &\qquad\qquad + \underline{E\left\{\mathbf{s}_{r,j}(n)\mathbf{s}_{d,j}^T(n)\right\}} \\ &\qquad\qquad \left.\left. + \underline{E\left\{\mathbf{s}_{r,j}(n)\mathbf{s}_{r,j}^T(n)\right\}}\right)\right)^{-1} \\ &\times \left(E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\} + \underline{E\left\{\mathbf{s}_{d,j}(n)s_{r,i}(n)\right\}}\right. \\ &\qquad \left. + \underline{E\left\{\mathbf{s}_{r,j}(n)s_{d,i}(n)\right\}} + \underline{E\left\{\mathbf{s}_{r,j}(n)s_{r,i}(n)\right\}}\right). \end{aligned} \quad (32)$$

With the assumptions given in Section IV-B that the reverberant and direct components are uncorrelated and that the reverberation is normally distributed noise with a mean of zero and is uncorrelated in each input channel,

the *underlined* terms in (32) can be removed. Furthermore, according to (29) we can replace $E\left\{s_{r,i}^2(n)\right\}$ with $\gamma_i E\left\{s_{d,i}^2(n)\right\}$, giving

$$\begin{aligned}\Psi_i =& 1 - E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\}^T \\ &\times \left(E\left\{s_{d,i}^2(n)\right\}(\gamma_i+1)\left(E\left\{\mathbf{s}_{d,j}(n)\mathbf{s}_{d,j}^T(n)\right\}\right)\right)^{-1} \\ &\times E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\}. \end{aligned} \tag{33}$$

4) The generalization in (25) allows (33) to be written as

$$\begin{aligned}\Psi_i =& 1 - E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\}^T \\ &\times \left(\sigma_{s_{d,i}}^2(\gamma_i+1)\sigma_{s_{d,j}}^2\right)^{-1} \\ &\times E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\} \\ =& 1 - \frac{E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\}^T E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\}}{\sigma_{s_{d,j}}^2\sigma_{s_{d,i}}^2(\gamma_i+1)}. \end{aligned} \tag{34}$$

5) By defining the cross-correlation coefficient vector between the direct-path vector $\mathbf{s}_{d,j}(n)$ and the direct-path sample $s_{d,i}(n)$ as

$$\mathbf{c}_{s_{d,i}s_{d,j}} = \frac{E\left\{\mathbf{s}_{d,j}(n)s_{d,i}(n)\right\}}{\sqrt{E\left\{s_{d,j}^2(n)\right\}E\left\{s_{d,i}^2(n)\right\}}} \tag{35}$$

the resulting solution for Ψ_i is found to be dependant on the direct-path IR correlation $\mathbf{c}_{s_{d,i}s_{d,j}}$ and the reverberant-to-direct level γ_i

$$\Psi_i = 1 - \frac{\left\|\mathbf{c}_{s_{d,i}s_{d,j}}\right\|^2}{\gamma_i+1}. \tag{36}$$

Combining (36) and (27), it is seen that the interchannel correlation $\mathbf{c}_{m_jm_i}$ is proportional to the correlation of the direct-path IRs and inversely proportional to the relative reverberant energy in the IR

$$\left\|\mathbf{c}_{m_jm_i}\right\|^2 = \frac{\left\|\mathbf{c}_{s_{d,i}s_{d,j}}\right\|^2}{\gamma_i+1}. \tag{37}$$

When the reverberation level is high, the denominator of (37) will dominate: the input cross-correlation will be low and the output error level large. This trend is visualized in Fig. 4, where it is seen that when the level of reverberation is 60 dB higher than the direct part, the correlation between the two microphones is approximately zero, irrespective of the correlation between the direct path sound. It can also be seen that when the direct sound level is 30 dB greater than the reverberant component level, the overall correlation between the two input signals is dominated by the correlation between the direct path signals $s_{d,i}$ and $s_{d,j}$. As typical concert-hall recordings of music are generally made with the microphone diaphragms within the reverberation radius [6], the most applicable part of Fig. 4 is to the left-hand side of the central line of the γ_i axis.

VI. Experimental Results

Two experiments are now reported which used empirical measurements made in a large reverberant space to investigate optimum filter length and to validate the signal model proposed in Section IV-A.

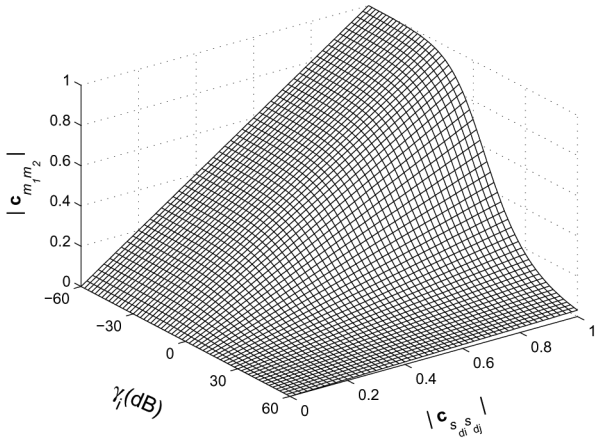

Fig. 4. Input signal correlation (max) $\mathbf{c}_{m_jm_i}$ as a function of direct-path correlation $\mathbf{c}_{s_{d,i}s_{d,j}}$ and the reverberant-to-direct ratio γ_i.

A. Method

To create the test stimulus, white noise was reproduced from a single loudspeaker on a stage in a 2000-m^3 concert hall (reverberation time $T_{60} = 1.8$ s at 1 kHz) and recorded using a pair of cardioid microphones. The state of one adaptive filter and the final NMSE level was noted after 30 s, which will be discussed shortly.

The frequency-domain NLMS algorithm parameters used for the analyses were as follows:

- overlapping factor: 4;
- delay of unfiltered channel: 500 samples.

B. Selection of Filter Length

The purpose of the new system is to remove the correlated direct-component from the input signals, so the length of the adaptive filter is designed to be the length of the direct-sound components (i.e., up to the mixing time of the intermicrophone impulse response, which is equal to L_r in Fig. 2). Kurtosis was used as a measure of normality, as the reverberant component of an impulse response can be defined as that part where the local distribution is normal (Gaussian) [17]

$$\text{kurtosis} = \frac{E\left\{x-\mu\right\}^4}{\sigma^4} \tag{38}$$

where μ is the mean and σ the standard deviation of x. A population of samples with a normal distribution has a kurtosis of 3. To empirically determine the onset of the reverberant component in the adaptive filter, a running measure of its kurtosis was calculated. The rectangular averaging window was 96 samples and the window was advanced by 12 samples between consecutive block averages. Fig. 5 shows the kurtosis of the adaptive filter for different filter taps, wherein it can be seen that a normal distribution occurs at about 1000 samples (23 ms), with two low-level reflections occurring later. This probably explains why no subjective improvement in system performance was achieved for increases in filter length over 1024 taps [13], and therefore the

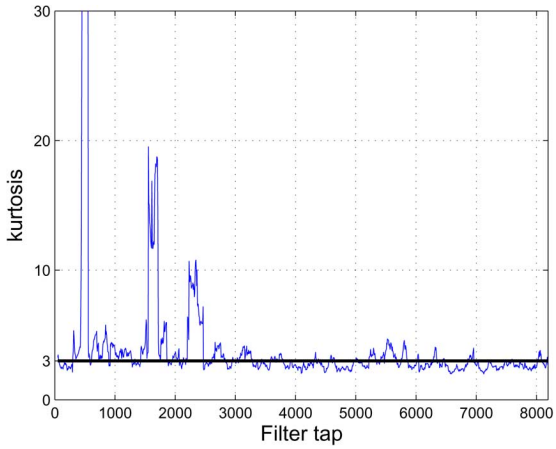

Fig. 5. Local kurtosis of adaptive filter for spaced microphone recording with on-axis loudspeaker reproducing white noise in a concert hall. Filter length was 8192 taps, the overlapping factor was 4.

length of the filter in the subjective evaluation of the upmixer reported shortly was also 1024.

C. Validation of System Model

In a second measurement, the spacing between the microphone diaphragms was varied while keeping the distance from the source constant (about 3.5 m). The exception to this was a case when the microphone pair was taken to the back of the hall, about 26 m from the loudspeaker source yet only 1 m apart. In the latter case, even though the correlation of the direct sound and early reflections may be high, the reverberant-to-direct level ratio (γ) would also be high. As predicted by (37) (summarized in Fig. 4); when γ is larger than about 15 dB, the overall interchannel correlation $\mathbf{c}_{m_j m_i}$ is dominated by γ and is no larger than 0.4, giving an NMSE of about -1.5 dB. In other words, when recordings made in a reverberant field are upmixed, the level of the output signals are similar in level to the input signals.

Looking at Fig. 6, the far-away microphone pair (i.e., case 1 m*) is highly correlated at low and high frequencies (about 0.95 at 100 Hz and 0.8 at 12 kHz), and therefore the NMSE is low (about $-$ 15 dB). At mid-frequencies, however, the signals are less correlated (0.38 at 1.5 kHz) giving a higher NMSE (close to 0 dB). A similar trend is also seen for the other microphone pairs. This can be explained by two factors: First, if the microphone diaphragm spacing is small compared with the wave-length λ, there is little decorrelation effect as the sound pressure is similar at each microphone [21]. This explains why for the 6 cm spacing the microphone signals are highly correlated up to about 1 kHz. Second, the reverberant-to-direct ratio reduces at high frequencies due to air absorption (about 0.1 dB per meter at 4 kHz but only 0.001 dB/m at 100 Hz; [23, p. 224]) and sound absorption from objects in the room such as soft chairs and carpets.

The relationship of microphone signal cross correlation and NMSE shown in Fig. 7 support the signal model. The model is less robust when the correlation is high, but it must be remembered that the theoretical derivation assumed stationary impulse response statistics as well as a noise-free operating environment;

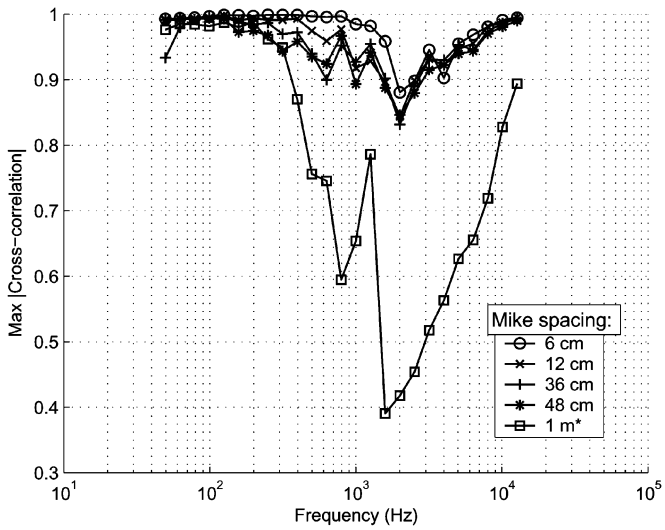

Fig. 6. Maximum absolute cross-correlation (in a 23-ms window, averaged over the recording) for different microphone spacings. White noise was reproduced with a loudspeaker in a 2000-m^3 concert hall and recorded with microphone pairs with a variety of spacing. The "1 m*" recording was with a microphone-spacing of 1 m, but was 26 m from the source; all other recordings were made 3 m from the source.

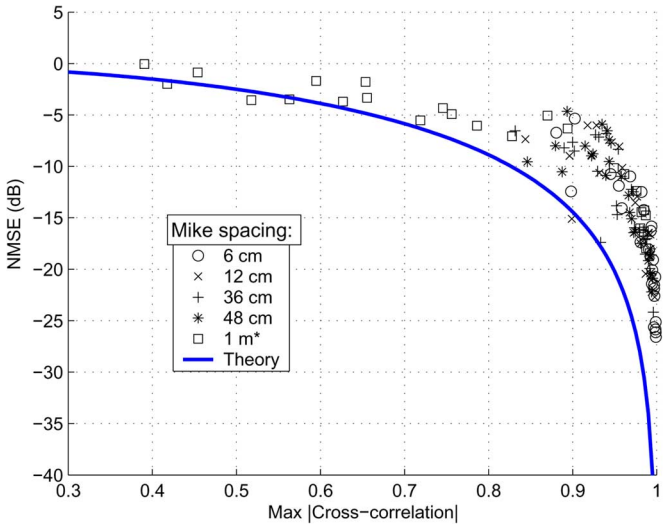

Fig. 7. Effect of intermicrophone correlation on NMSE, using the data presented in Fig. 6. Different markers represent NMSE and cross correlation of input signals for different frequencies (i.e., the same data as was used in Fig. 6).

both of which were not the case (due to air turbulence and background noise). Also, because the filter update step-size was finite, the optimal solution could never be exactly met and this results in gradient noise [9] which would limit the minimum level of the error signal. Other studies [25], [26] have remarked how it is very difficult to get NMSE statistics for practical adaptive filtering applications (such as echo canceling or dereverberation) less than about -20 dB.

VII. Subjective Evaluation

A subjective evaluation of the new ambiance extractor upmix system was undertaken to investigate if people preferred upmixed "surround-sound" loudspeaker audio scenes or the original "legacy stereo" two loudspeaker scene.

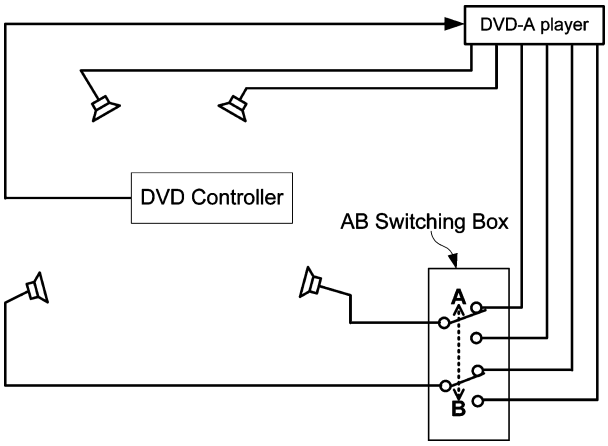

Fig. 8. Signal schematic for preference experiment. All audio outputs were from a DVD-A player (the low-frequency effect channel was a normal, full-bandwidth signal). The rear-loudspeaker channel switching device was a passive switching box with an "A–B" clicking switch. The channel gain-trims are not shown.

A. Method and Stimuli

This experiment was conducted in an acoustically treated room approximately 50 m^3 with an RT_{60} of approximately 0.5 s. Four loudspeakers[3] were used, arranged in the conventional 2/2 ITU-775 [27] configuration (with no center speaker), with front loudspeakers at $\pm 30^{\circ}$ and rear loudspeakers at $\pm 120^{\circ}$. The listener sat on a nonrotating chair at the sweet-spot, 2.3 m from each loudspeaker. The loudspeakers were calibrated so as to produce an equal sound pressure level at the listening position (74$\pm$0.5 dB, unweighted, slow time averaging, using pink noise). The stimuli were burned onto a DVD disc (recorded at 44.1 kHz, 16 bit).

The method of paired comparison was used to evaluate the new upmixer in terms of overall preference [28]. For each trial the subject was presented two stimuli A or B, which the subject could freely switch between using a 4-in, 2-out audio signal switching box, as shown in Fig. 8. Stimulus A or B corresponded to one of four scenes; a variant of the new upmixer or the 2/0 scene. Using a computer graphical user interface, the subject reported which scene was preferred.

Two groups of people undertook the experiment: five audio engineers and 11 musicians. The engineers were all past sound recording (Tonmeister) students, each with at least three years of experience with sound recording practice. The musicians (most of whom are professional) were enroled on an intensive music performance or composition program.

The original stimuli were created by reproducing an anechoic recording of a sung solo voice and solo viola with a loudspeaker in a concert hall (as in Section VI-A) and recording it with a pair of microphones.[4] This approach was chosen over using a conventional recording as it is more representative of a real musical instrument performance in a hall in terms of the auditioned sound qualities, yet allows measurements to be repeated much more consistently than with a "live" musical performance by a musician. The loudspeaker was either equidistant to each microphone ("on-axis"), or 3 m off-axis. There were three variants of the new upmix system which were compared with each other and with a reference 2/0 scene created by reproducing the original two-channel recording with just the front loudspeaker pair. Thus, the enhancement of the new upmix system compared with conventional two-loudspeaker presentation could be evaluated. The three upmixer variants, plus the 2/0 scene, are summarized as follows:

1) unmodified 2/2 upmixer (as described by Fig. 1);
2) 2/0 (only the front two loudspeakers were active);
3) 2/2 upmixed sound scene with rear loudspeaker channels delayed by 10 ms;
4) 2/2 upmixed sound scene with rear loudspeaker channels attenuated by 6 dB.

The front loudspeaker signals were delayed by the 500 sample delay D, plus 1024 samples to account for the delay of the upmix signal processing system. The 10-ms delay was chosen as sound radiated from two spaced loudspeakers, with one delayed by this amount increases perceived listener sound envelopment [19], yet with music and speech audio signals a listener generally hears a single sound image localized in the direction of the nondelayed loudspeaker, in accordance with the precedence effect [29]. The 6-dB attenuation was chosen by informal empirical perceptual evaluation with the recorded music, as this level gave a natural-sounding balance of reverberation around the listener [13].

Each of the four recordings were presented with each of the four scene configurations. Therefore, there were six pair-wise comparisons for each of the four fragments, giving 24 unique paired comparisons. This was presented twice to each subject, with the A–B stimuli order reversed for the second presentation. Each subject was presented the 24 excerpts of music twice, with a 5–20-min break in between. The subjects were asked: "Which sound scene do you prefer: A or B?" Once they selected either option, a pop-up window prompted them to confirm and advance the DVD to the next track. The subjects were told they should think about the preference task as if they were evaluating a product which they might purchase for home entertainment.

B. Results of Preference Experiment

Results for the paired comparisons are shown in Fig. 9 which shows how often a particular stimulus was preferred over another. The data is split into the musician and engineer group. The total number of trials ($\mathrm{nTrials}$) was 24 scene pairs $\times$ 2 runs $\times$ (the number of subjects), which was 240 for the engineer group and 480 for the musician group; 95% confidence intervals ($\pm 2\sigma$) were calculated as follows:

$$\sigma = \sqrt{\mathrm{nTrials} \times P(A)(1 - P(A))} \tag{39}$$

where $P(A)$ is the probability of a subject picking a scene configuration A by chance (which was 0.25 as there were four scenes).

C. Discussion

From the results of the preference choice analyses shown in Fig. 9, it can be seen that the standard 2/0 two-loudspeaker audio scenes are consistently dispreferred over 2/2 upmixed scenes for

[3]Type 1031 manufactured by Genelec.

[4]Examples of the audio stimuli can be found online at http://www.JAR-lab.com/upmixer.

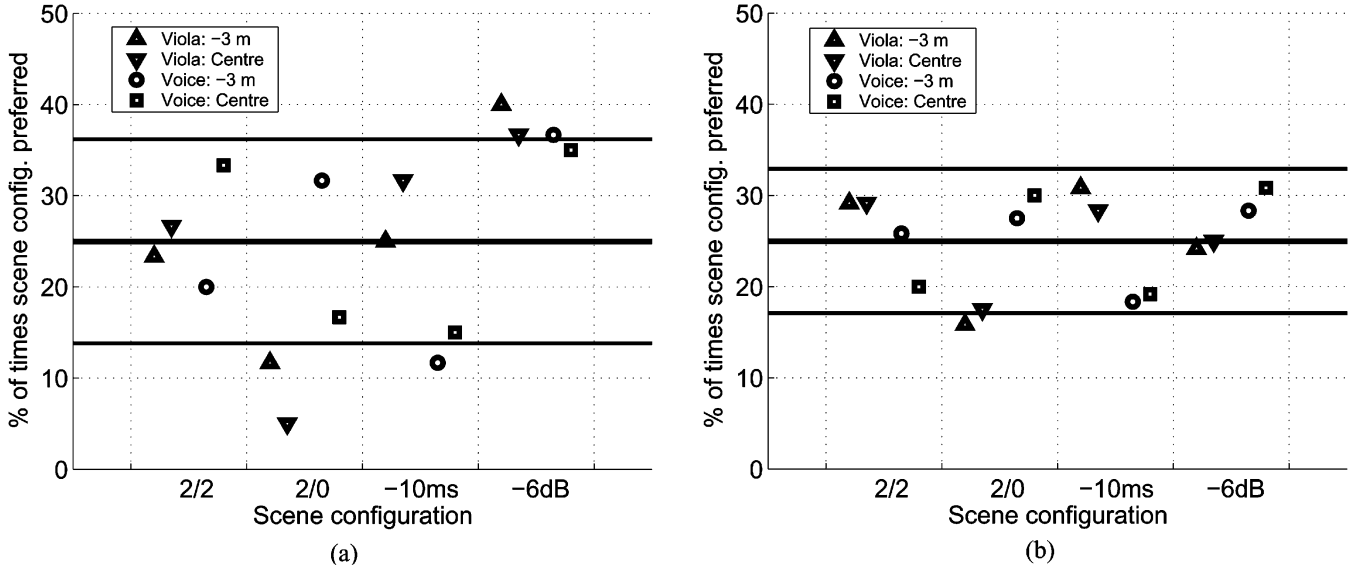

Fig. 9. Average preference of one of the four audio scene configurations over all remaining three. Scenes were presented as a paired AB comparison and results are grouped by scene configuration. All scenes were the upmixed 2/2 scene, except scene 2/0 (which was just the front loudspeaker pair; i.e., "legacy stereo"). Central solid line shows likelihood of preferring a scene by chance (25%, i.e., if the subjects randomly pressed A or B) and flanking lines are 95% confidence intervals. If the marker is above the upper-most line, then this scene configuration was preferred significantly more than the others for the particular audio sample. If the marker is between the upper and lower lines; this scene was neither preferred nor dispreferred, and if it is below all lines then this scene was preferred less than the other scenes. There were four stimuli and two presentations; so 240 responses for the audio engineer group and 480 for the musician group. (a) Audio engineer group (five subjects). (b) Musician group (ten subjects).

the viola recordings by both the engineer and musician subject groups.

The upmixed scenes with attenuated rear channels were significantly preferred over all other scenes for the engineer group (though this trend was less strong for the voice stimuli).

Both listening groups generally preferred the 2/2 scene with the 10-ms delay *less* than the other scenes for the voice stimuli. This might be because the 10-ms delay destroyed the pair-wise amplitude panning between the (correlated) reverberation components in the front and rear loudspeakers, with the reverberance image collapsing to the front speakers due to the precedence effect. The transient components of the voice may reveal this easier than the smoother temporal envelope of the violin. From a cross-correlation analysis of the side loudspeaker channels, it was found that these side channels (e.g., front-right loudspeaker signal and rear-right ambient signal) had a nonzero correlation. Together with a graphical image description (reported in [13]), this supports the idea that side reverberance images are created in upmixed sound scenes when there is no additional delay between the front and rear loudspeakers.

VIII. Conclusion

A new system has been proposed for enhancement of the spatial sound quality of a listening experience created with an unencoded two-channel ("stereo") audio signal. The system involves extraction of uncorrelated reverberation from sound recordings which can be reproduced with surrounding loudspeakers such as those in surround-sound home-theater or automotive audio systems. Sound components correlated within approximately 20 ms are removed by equalizing the input signals with respect to both magnitude and phase and then differencing the signals. A model for the automatic "ambiance extraction" system was proposed and validated with empirical measurements in a concert hall. The model predicts that the level of the extracted reverberation is related to the interchannel signal correlation, which in turn is related to both the correlation between the early-arriving sound to each microphone, and the ratio of the reverberant component of the room impulse response to the early-arriving component. Formal subjective evaluation of the system shows promising results to support the new upmixer as an effective means for enhancing spatial sound quality of reproduced audio scenes.

Acknowledgment

Author J. Usher would like to thank his principle doctoral advisors Dr. W. L. Martens and Dr. W. Woszczyk. The authors would also like to thank all those who took part in the listening tests from the Music and Sound Program at the Banff Centre, Banff, AB, Canada, and three anonymous reviewers for their helpful suggestions and comments.

References

[1] C. Avendano and J.-M. Jot, "Ambience extraction and synthesis from stereo signals for multichannel audio upmix," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, 2002, pp. 1957–1960.

[2] Y. Choi, S. Han, D. Lee, and K. Sung, "A new digital surround processing system for general A/V sources," *IEEE Trans. Consum. Electron.*, vol. 41, no. 4, pp. 1174–1180, Nov. 1995.

[3] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *J. Audio Eng. Soc.*, vol. 50, no. 11, pp. 914–926, 2002.

[4] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051–1064, 2006.

[5] P. M. Clarkson, J. Mourjopoulos, and J. K. Hammond, "Spectral, phase and transient equalization of audio systems," *J. Audio Eng. Soc.*, vol. 33, no. 3, pp. 127–132, 1985.

[6] A. Fukada, K. Tsujimoto, and S. Akita, "Microphone techniques for ambient sound on a music recording," in *Proc. AES 103rd Int. Conv.*, New York, 1997, paper 4540.

[7] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.
[8] S. Haykin, *Adaptive Filter Theory*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall, 2001.
[9] B. Widrow and J. M. McCool, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, no. 8, pp. 1151–1162, Aug. 1976.
[10] D. T. M. Slock, "On the convergence behaviour of the LMS and the normalized LMS algorithms," *IEEE Trans. Signal Process.*, vol. 41, no. 9, pp. 2811–2825, Sep. 1993.
[11] P. C. W. Sommen, P. J. VanGerwen, H. J. Kotmans, and A. J. E. M. Janssen, "Convergence analysis of a frequency-domain adaptive filter with exponential power averaging and generalized window function," *IEEE Trans. Circuits Syst.*, vol. CAS-34, no. 7, pp. 788–798, Jul. 1987.
[12] J. Benesty, "General derivation of frequency-domain adaptive filtering," in *Advances in Network and Acoustic Echo Cancellation*, J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, Eds. New York: Springer, 2001.
[13] J. S. Usher, "Subjective evaluation and electroacoustic theoretical validation of a new audio upmixer," Ph.D. dissertation, Schulich School Music, McGill Univ., Montréal, QC, Canada, 2006.
[14] V. Pulkki and T. Hirvonen, "Localization of virtual sources in multichannel audio reproduction," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 105–119, Jan. 2005.
[15] J.-M. Jot and A. Chaigne, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Proc. AES 103rd Int. Conv.*, New York, 1997, paper 4629.
[16] L. L. Beranek, *Concert and Opera Halls: How They Sound*. Woodbury, NY: Acoust. Soc. Amer. through AIP Press, 1996.
[17] M. R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *J. Audio Eng. Soc.*, vol. 35, no. 5, pp. 299–305, 1987.
[18] B. Blesser, "Interdisciplinary synthesis of reverberation viewpoint," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 867–903, 2001.
[19] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *J. Sound Vibration*, vol. 77, pp. 211–232, 1981.
[20] M. R. Schroeder and H. Kuttruff, "On frequency response curves in rooms," *J. Acoust. Soc. Amer.*, vol. 34, pp. 76–80, 1962.
[21] L. Cremer and H. A. Müller, *Principles and Applications of Room Acoustics*. New York: Applied Science, 1982, vol. 2.
[22] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 168–172, Mar. 2000.
[23] L. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 4th ed. New York: Wiley, 1999.
[24] H. Kutruff, *Room Acoustics*, 3rd ed. Essex, U.K.: Elsevier, 1991.
[25] G. W. Elko, E. Diethorn, and T. Gaensler, "Room impulse response variation due to temperature fluctuations and it's impact on acoustic echo cancellation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Kyoto, Japan, 2003, pp. 67–69.
[26] P. Hatziantoniou and J. Mourjopoulos, "Errors in real-time room acoustics dereverberation," *J. Acoust. Soc. Amer.*, vol. 52, no. 9, pp. 883–899, 2004.
[27] "Multichannel stereophonic sound system with and without accompanying picture," Int. Telecomm. Union Radiocommun. Assembly, 1994, ITU-R BS 775-1, Rec. BS.775-1, 1992–1994.
[28] "Sound system equipment-part 13: Listening tests on loudspeakers Int. Electrotech. Commission, Geneva, Switzerland, 1985, IEC 268-13.
[29] R. K. Clifton, "Breakdown of echo suppression in the precedence effect," *J. Acoust. Soc. Amer.*, vol. 82, no. 5, pp. 1834–1835, 1987.

John Usher (M'07) was born in Devon, U.K., on July 5, 1979. He received the B.Eng. degree in electroacoustics from the School of Acoustics, University of Salford, Salford, U.K, in 2001 and the Ph.D. degree from McGill University, Montréal, QC, Canada, where he was supervised by Dr. W. L. Martens; Prof. W. Woszczyk, and Prof. A. Bregman. His Ph.D. thesis was on the theoretical development and subjective evaluation of a new ambiance extraction upmixer.

Since 2006, he has been a Chief System Architect at Personics, Montreal, QC, Canada, a new high-tech earphone company, where he has filed over 15 patents. During his undergraduate program he worked with Bang and Olufsen for a year in Denmark, developing a multichannel audio upmixer for line-array loudspeakers, which resulted in a patent.

Dr. Usher is a member of the Audio Engineering Society, the Acoustical Society of America, and the National Hearing Conservation Association.

Jacob Benesty (SM'04) was born in 1963. He received the M.S. degree in microwaves from Pierre and Marie Curie University, Paris, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, Paris, in April 1991.

From November 1989 to April 1991, he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecommunications (CNET), Paris. From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ. In May 2003, he joined the University of Quebec, INRS-EMT, Montréal, QC, Canada, as an Associate Professor. His research interests are in signal processing, acoustic signal processing, and multimedia communications. He was a member of the Editorial Board of the *EURASIP Journal on Applied Signal Processing*. He coauthored *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006) and *Advances in Network and Acoustic Echo Cancellation* (Springer-Verlag, 2001). He is also a coeditor/coauthor of *Speech Enhancement* (Spinger-Verlag, 2005), *Audio Signal Processing for Next Generation Multimedia communication Systems* (Kluwer, 2004), *Adaptive Signal Processing: Applications to Real-World Problems* (Springer-Verlag, 2003), and *Acoustic Signal Processing for Telecommunication* (Kluwer, 2000).

Dr. Benesty received the 2001 Best Paper Award from the IEEE Signal Processing Society. He was the Co-Chair of the 1999 International Workshop on Acoustic Echo and Noise Control.