Audio Engineering Society

Convention Paper

Presented at the 121st International Convention
2006 October 5–8 San Francisco

A new upmixer for enhancement of reverberance imagery in multichannel loudspeaker audio scenes

John Usher

Department of Sound Recording, McGill University, Montreal, Canada

Correspondence should be addressed to (`jusher@po-box.mcgill.ca`)

ABSTRACT

This paper introduces a new signal processing system which enhances reverberance imagery (i.e. perceived ambiance or listener envelopment) in loudspeaker audio scenes. Sound components which affect reverberance imagery are extracted from a pair of unencoded audio signals and are radiated with two additional loudspeakers behind the listener. The new "ambiance extraction" system improves upon all extant systems by using a novel automatic (blind) equalizer based on the normalized least means square (NLMS) algorithm to align the input signals with respect to both level and time in order to create the difference signal. The alignment is typically undertaken using a 1024-tap frequency and ±10 ms time equalizer, which allows sound components with a high short-term correlation to be removed from the input audio signals. Subjective and objective evaluation was undertaken with recordings of solo musical performances in a concert hall, and show that the new system provides a computationally practical, high-quality solution to the problem of ambiance extraction for audio upmixing.

1. INTRODUCTION

Multichannel loudspeaker audio is a ubiquitous phenomenon in home theatre and automotive entertainment systems throughout the world today. An upmixing device that can create from our blooming two-channel music collections an additional set of signals which can be reproduced with surrounding loudspeakers is therefore more relevant than ever. The benefits of reproducing recorded music in "surround sound" are two-fold: Allowing a more *immersive* and *natural* sounding impression of the recording environment with an increased sense of *presence* of the musician, and with secondary effects of making it easier to discern spatial nuances in the musical performance and reveal subtle aspects of the recorded sound which would be otherwise hidden with two-loudspeaker reproduction; *"A good [audio] system not only allows listeners to hear sources at different directions, but also improves a listeners' ability to understand simultaneous sources and monitor a complex auditory scene"* [1].

This paper is about such an audio upmixer and addresses the major shortcomings of existing upmixers. These shortcomings have been previously ignored due to certain assumptions about the music intended to be upmixed; that the sound recordings are created with intensity-panned audio mixes. The new upmix system is designed for use with both these recordings and recordings made using multiple spaced microphones; a method generally employed for recording concert-hall musical performances where *time-delay panning* [2] occurs due to the sound sources being closer to some microphones than others.

It was shown in electronic measurements reported in a companion paper [3] that music recorded with spaced microphone pairs could not be processed by extant upmixers as effectively as when intensity-panned recordings are used. This conclusion was qualified by an analysis of signal correlation between rear and front channels for recordings made using spaced microphones and off-axis sources (i.e. where time-delay panning occurred). It was found that when time-delay panned recordings were used the extant upmixers investigated[1] had undesirable processing artifacts such as rear loudspeaker signals which were highly correlated with both front loudspeaker signals. It was argued (though not subjectively tested) that if the rear signals are correlated with both the front loudspeaker signals (within a window of about 10 ms), then the upmixing process could distort the frontal (source) image by pulling it towards the rear loudspeakers. Furthermore, the output signals of these upmixers often exhibited undesirable level "pumping" as the upmixer tries to compensate for the increase in rear channel level.

It will be shown that time-alignment of the input signals is possible with the new upmix system presented here, and that this firstly ensures minimal distortion of source imagery by the upmixing process, and secondarily creates new reverberance images with a high image strength at lateral directions. The new upmixer is designed so as to provide a consistent audio quality for a variety of recording methods (i.e. those with both time and amplitude panning) using a novel implementation of the NLMS adaptive signal processing algorithm.

[1] Circle Surround II and Dolby Pro Logic II were investigated.

1.1. Source and reverberance imagery in reproduced sound scenes

Studies of sound imagery in spatial (multichannel) audio systems have shown that there are two aspects of auditory spatial imagery which have meaningful perceptual relevance to listeners, and significantly influence ratings of both the overall sound quality [4] and preference [5]: the perception of sound relating to aspects of the recorded sound source (*source imagery*; of which *Auditory Source Width*- ASW- is a component) and sound relating to aspects of the recording environment (*reverberance imagery*; of which *Listener Envelopment*- LEV- and *spaciousness* is a component).

The temporal factors affecting the distinction between source and reverberance imagery are thought to be the same principles relating to the *precedence effect*; whereby early arriving, high level sound (direct sound and early reflections) primarily affect the source image [6] and later arriving low level sound (reverberation) affect the reverberance image [7].

1.2. Interaction of source and reverberance imagery in loudspeaker audio

In order to provide a sensitive means for reporting auditory spatial imagery in multichannel loudspeaker audio scenes, a new computer-driven graphical mapping system was developed [8, 9]. This enabled the spatial geometry of auditory source and reverberance images to be visualized with a two-dimensional drawing. Ellipses drawn with the GUI to represent the location and spatial extent of the auditory image in the horizontal plane could be analyzed to measure reported image width, distance and range [9, 10]. A subjective experiment using the GUI was undertaken to investigate the spatial interaction of source and reverberance images in multichannel loudspeaker audio scenes [11].

The two main findings were:

- Control of *reverberance* image direction and width involves similar amplitude panning principles as for *source* images. Pair-wise panning of source images around the listener using loudspeakers has been investigated extensively [2, 12, 13, 14], and the findings in this earlier study [11] support the hypothesis that a generalization of these panning principles is, at least to some extent, possible for reverberance images. More specifically,

it was found that reverberance images are localized to the side of the listener by radiating correlated recorded reverberation to side loudspeakers at +30° and 120°.

• The perceived width and direction of a source image panned at 0° azimuth was significantly affected (i.e. distorted) by the panned direction of a reverberance image. This *spatial distortion* was reduced as the perceived separation between source and reverberance images was increased. These findings support the analogy between source and reverberance image interaction and a target/ masker paradigm; whereby a source (target) image can be unmasked to increase understanding (or "readability" [15]) of the sound scene by creating a reverberance (masking) image which seems to originate from a different or distributed direction.

These findings led to three general design criteria and specific subjective and electroacoustic evaluation methods for a high quality upmixer to enhance the perceived quality of reverberance imagery, which were discussed in a previous paper [3] and are summarized shortly.

1.3. Extant upmixing systems for enhancement of reverberance (ambiance) imagery in audio

The new upmixer is best described as an *(unsupervised) ambiance extraction system* [16], in contrast to "repanning" systems such as "Trifield" [17] which aim to affect primarily the *source* image components by re-radiating the sound with at least three frontal loudspeakers. The new upmixer would therefore compliment such repanning systems if, for example, the listener or system designer wanted to use all five loudspeakers as with a conventional 3/2 ("5.1") loudspeaker system.

The word "unsupervised" (or "blind") is often used with signal processing techniques where nothing is assumed a priori about the audio signals themselves. Linear matrix converters [18, 16] are also upmix systems, but generally assume *encoded* signals (such as the classic Scheiber encoder/ decoder [19] - the principle behind Dolby Surround).

Blind upmixers differ from linear matrix converters in two important ways:

- Blind upmixers are active: the data processing structure is dependant on the particular input signal properties and the input-to-output relationship will be different for different input signals. For example, the scaling parameters for the input signals to derive the new signals (via addition and subtraction) are adaptive and dependant on the particular audio input signals.
- In blind upmixers the input signals do not have to be specially encoded. These upmix systems are intended to be used with conventional "off-the-shelf" two channel recordings and it is assumed the sound engineer who mixed the CD intended it to be reproduced with a conventional two loudspeaker arrangement (i.e. a discrete 2/0 loudspeaker system).

The "ambiance extraction" process of extant blind audio upmixers rely on a common fundamental assumption: that a single dominant image direction exists at a given time and has been created using amplitude-panning techniques. The methods for finding the principal image direction vary; e.g. by an estimate of the level of each input channel (such as with Logic7 [20]), or using a bootstrapped mechanism which aligns the magnitude of the input signals using a comparator to minimize the difference signal (such as Dolby Pro Logic II [21] or the Aarts/ Irwan system [22]).

No existing system has a mechanism to deal with input signals where the direct sound component arrives in one channel before the other (such as occurs with time-delay panning). Furthermore, the level alignment procedure is generally a global level adjustment rather than a detailed frequency-dependent gain (an exception is the Avendano/ Jot system [16], which can discriminate between correlated and uncorrelated sound components as a function of frequency, but can not discriminate between correlated and uncorrelated components within the same band).

2. SUBJECTIVE AND ELECTRONIC DESIGN CRITERIA FOR A NEW UPMIXER

From a review of the literature relating to "spatial release from masking" [23, 24, 25] and the results of the experiment on source and reverberance image interaction discussed earlier [11], a set of design criteria for the new upmixer were proposed (explained more thoroughly in; [3]). The three criteria are now summarized as they relate to both a subjective and electroacoustic evaluation:

1. *Spatial distortion of the source image in the upmixed audio scene should be minimized.*

The principal aim is to maintain a similar spatial sound character of the source image in both the upmixed and reference 2/0 audio scenes (i.e. a high *fidelity* of the source image). It was proposed that if those sound components that affect only the reverberance image could be electronically extracted, then these could be radiated from loudspeakers behind the listener to create new side reverberance virtual images; leaving the source image spatially undistorted. The efficacy of the upmixer for satisfying this can be measured using the graphical mapping system to compare elicited source image geometry in the upmixed and reference 2/0 scenes (e.g. a comparison of source image width and direction in the original 2/0 and upmixed 2/2 scene, as indicated with the graphical response technique).

As mentioned, it is the short-term correlated sound components in a pair of audio signal which affect source imagery (i.e. when these signals are radiated with a loudspeaker pair), so if just one of the rear loudspeaker signals is uncorrelated with one of the front loudspeaker signals, then the source image should not be affected. A way of measuring this in electronic terms is to ensure that the four loudspeaker signals are *diagonally uncorrelated*. As will be described shortly, loudspeaker signals on the same side (e.g. signals feeding the front-right and rear-right loudspeakers) should have some degree of correlation so that side reverberance images can be formed by coherent pair-wise panning [26].

To maximize the source image fidelity in the upmixed audio scene, source image components S_L and S_R (see figure 5) should not be radiated from the rear loudspeakers in the upmixed sound scene. If they were, then they could perceptually interact with the source image components radiated from the front loudspeakers and cause the source image to be distorted (as was shown in a previous study [11], side phantom images can be created if correlated audio signals are radiated from front and rear loudspeakers). Therefore, all those sound components which contribute to the formation of a source image should be removed from the rear loudspeaker signals, yet those source image components radiated from the front loudspeakers should be maintained. A way of measuring this in electronic terms is to ensure that the signal RS is uncorrelated with signal L, and that LS is uncorrelated with R. For a signal sampled at time n, this is summarized expressed in (1):

$$0 \approx \sum_{n=-\infty}^{\infty} RS(n)L(n-k) \quad \text{and} \quad 0 \approx \sum_{n=-\infty}^{\infty} LS(n)R(n-k). \quad k = \pm 0, \pm 1, \pm 2, \ldots, \pm N. \tag{1}$$

The lag range N should be equal to 10-20 ms (500-1000 samples for a 44.1 kHz sample-rate digital system), as it is the early sound after the direct-path sound which primarily contributes to spatial aspects of source imagery (such as source width) and the later-arriving sound which affects reverberance imagery [7, 27, 28].

2. *Reverberance imagery should have a homogenous distribution in the horizontal plane; in particular, reverberance image directional strength should be high from lateral (±90°) directions.*

As was shown by Hiyama [29], a perceptually horizontally isotropic reverberant sound field (PHIRF)[2] can be achieved to quite a reasonable degree using just the four loudspeakers in the 2/2 ITU-775 configuration, or by reproducing four plane waves of recorded reverberation with a wave field synthesis system [31]. However, when reverberation is reproduced from lateral directions (±90° azimuth), the sense of "spatial impression" [27] or LEV [28] is stronger.

The implication of this is that in order to create new reverberance images to the side of the listener, the side loudspeaker channels (e.g. R and RS) should have some degree of correlation. Under such circumstances, pair-wise amplitude panning could occur between the two loudspeakers; with the perceptual consequence that the reverberance image would be pulled *away* from the side loudspeakers and to a region *between* them (as found in the previous study [11]). This is mathematically ex-

[2] A PHIRF means a *homogenous* reverberance image "one in which no direction is preferentially treated" [30].

pressed in (2):

$$0 \neq \sum_{n=-\infty}^{\infty} LS(n)L(n-k) \quad \text{and} \quad 0 \neq \sum_{n=-\infty}^{\infty} RS(n)R(n-k), \quad k = \pm 0, \pm 1, \pm 2, \ldots, \pm N. \tag{2}$$

Again, N would be equal to 10-20 ms.

3. *The new system should not be dis-preferred to a conventional 2/0 system.*
   In the context of a home musical listening experience (i.e. listening for *pleasure*- not as part of a critical listening experience) the new system should be preferred over a reference 2/0 reproduction created using the same recording. The listening tests were actually undertaken in a laboratory setting, but the listeners were asked to imagine they were listening for pleasure in the preference experiments.

3. THEORY AND DESIGN ON THE NEW UPMIXER

An overview of the method for up-mixing two input audio signals is summarized in figures 6 and 7, and has four important components:

1. Filtering a first input audio signal with respect to a set of filtering coefficients (typically, with a 1024-tap FIR filter).
2. Time-shifting a second audio signals with respect to the first signal (typically with a delay of about 5 ms).
3. Determining a first difference between the filtered and the time-shifted signals. This difference signal is then radiated with a separate loudspeaker.
4. Adjusting the set of filtering coefficients based on the first difference so that the difference signal is essentially orthogonal to the first input signal.

So as to reflect the artistic mixing intentions of the audio engineer/ producer who mixed the musical recording to be upmixed (i.e. the first design criteria), the new upmixer is developed as a *natural spatialization algorithm*. Rumsey distinguishes this class of upmixers from more "artificial" approaches: *"... [natural spatialization algorithms] are not "effects" which explicitly add reverberation to simulate acoustic spaces other than those implied in the original program material, but rather they attempt to work on the spatial information already present in the original material"* [32]. This approach to upmixing ensures that if there is no spatial information in the recording (i.e. no recorded or artificial sound reflections/ reverberation), then no "ambiance" should be extracted by the upmixer. Strictly speaking, according to this maxim an effective natural spatialization algorithm should be able to extract the ambiance even from a mono recording. Such an algorithm would involve blind source-separation (blind-deconvolution) techniques [33], and doubtlessly these methods will be applied to audio upmixing in the near-future. However, as is shown in the thesis [34], the level of the extracted ambiance signal is actually related not to the level of recorded reverberation in the original material, but to the degree of correlation between the stereo input signal.

A very important feature of the NLMS algorithm, which makes it ideal for satisfying the design criteria, is the self-orthogonalizing properties of the system. According to the principle of orthogonality [35] (see appendix 3.1), when the adaptive filter has converged to its optimum solution (i.e. when the rate if change of filter coefficients is minimum), then the filtered output error signal (e.g. the signal feeding the right-surround loudspeaker) is uncorrelated with the input signal which was filtered (which is the diagonally opposite loudspeaker signal- which would be the front-left signal with the previous example). This elegant relationship was empirically verified using a variety of test signals [34] and was found to be accurate for stationary signals such as noise, and reasonably matched for coloured signals such as viola and voice (the correlation was about 0.35 maximum).

4. ELECTRONIC PERFORMANCE EVALUATION

4.1. Spectral and temporal alignment of input signals

A study into the adaptive filter conditions for different sound source locations for a two-microphone recording shows how the adaptive filter can align the input signals in both the time and frequency domain.

The recordings were made by reproducing a white noise signal with a loudspeaker in a 2000 m^3 concert hall (Pollack Hall at McGill University) and recording them with various microphone-pair configurations: an AB microphone pair which was spaced 50 cm; an XY coincident pair; and an ORTF pair, with diaphragms 16.5 cm apart and angled at 110° (all microphones were cardioid, B&K type 4011). The recordings were made with the loudspeaker located at a number of positions relative to the central "on-axis" location (i.e. the axis equidistant to the each microphone).

Looking at the adaptive filters in figure 1, it can be seen that as the loudspeaker source is moved from the -3 m to the 3 m position the peak grows from about 0.3 V to 1.0 V (for the ORTF arrangement); a change of over 10 dB. This is expected as the source is moving closer to the right-hand microphone, so the left-hand channel must be boosted to compensate to cancel the direct sound components. (The steering servo in the Dolby Pro-Logic II system responds in a similar way by boosting the lower-level channel before the difference signal is calculated [21].) With the new upmixer, this gain is applied on a frequency-by-frequency basis as if there is an N-band "equalizer" (here, N=1024).

Besides aligning the two input signals spectrally, the adaptive filter also aligns them temporally. This can be seen in the time-domain filter response shown in figure 1. For the coincident pair (XY) configuration, there is no time shifting of the main peak of the adaptive filter when the sound source moves from the 3 m to the -3 m position. This is because there is no change in time-of-arrival change due to the coincident diaphragms. The largest change is observed with the AB configuration (about 100 samples), which is expected as this has the largest diaphragm spacing. Details of the filter convergence are given in the thesis [34], where it is shown that the filter converges to the optimal solution within 0.1 seconds from initialization.

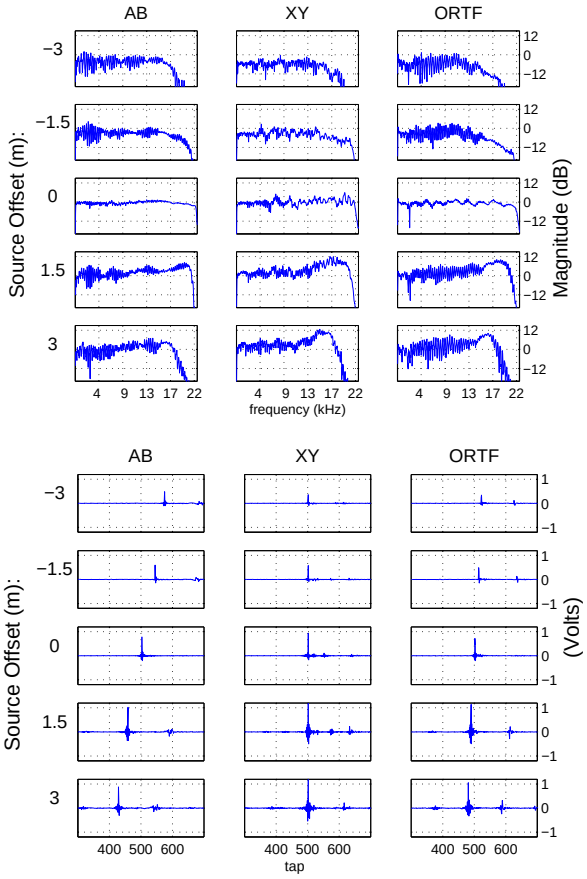


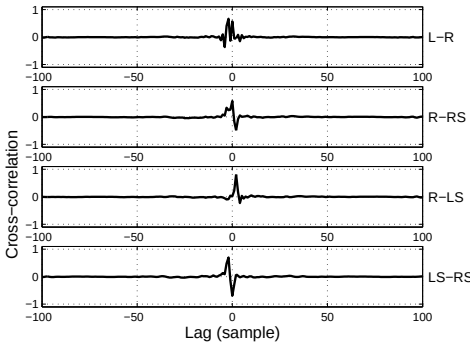
Fig. 1: Time and frequency domain representation of one adaptive filter for a white-noise sound source reproduced by a single loudspeaker in a 2000 m^3 concert hall, at five different locations. The 0 m offset location is when the sound source was equidistant to each microphone. The loudspeaker source was moved ±3 m from the central 0 m location. The AB microphone pair was spaced 50 cm, the XY was a coincident pair, and the ORTF was spaced 16.5 cm and angled at 110° (all microphones were cardioid, B&K type 4011). The input signal delay (delay 1 in figure 7) was 500 samples.

4.2. Correlation analysis of output signals

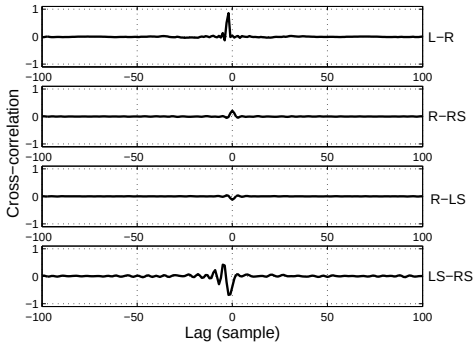
With the noise recordings used in the filter analysis, the electronic correlation between all output signals of the upmixer was measured (averaged over a 10 second portion). This was compared with the output response for the same test material using a popular commercial audio upmixer,[3] as summarized in figure 2.

The time-averaged cross-correlation measurement undertaken is not ideal, as temporal variation in the **IACC** has been shown to affect the width of source images [36]. However, it gives a basic insight into how the output channels of the upmixers may be be perceived. As mentioned in the subjective design criteria: *Spatial distortion of source image (compared with 2/0 loudspeaker audition) should be minimized.* In the electronic design criteria this was translated as: *Signal RS must be uncorrelated with signal L, and LS uncorrelated with R.* This was based on the assumption that it is only those correlated sound components between the mike channels which contribute to spatial properties of the source image. So looking at the correlation between signals *R* and *LS* in figure 2, it is seen that these signals are quite correlated for the commercial upmixer, yet these signals are uncorrelated for the new upmix system; in accordance with the design criterium.

Output signal correlation



(a) Commercial upmixer



(b) New upmixer

Fig. 2: Cross-correlation coefficient between input and output signals of a commercial upmixer and the new system (e.g. the third subplot R-LS shows the cross-correlation between diagonally opposite signals feeding the front-right and rear-left loudspeakers). The test audio signals were from a two-mike recording of white noise reproduced with a loudspeaker in concert hall. The noise source was slightly off-axis, hence the main peak occurs at a lag of -4 samples.

[3] Dolby Pro-Logic II, on a Marantz model SR4400 "AV Surround Receiver".

5. SUBJECTIVE EVALUATION

5.1. Image-mapping experiment

Graphical mapping of imagery in reproduced sound scenes provides a direct way to describe the spatial envelope of auditory images [9, 37, 38]. With such a sound character description technique, the listener is provided with a top-down view of the listening environment and is asked to represent the perceived location of the auditory images. This description encompasses a two-dimensional plan of the imagery in the horizontal plane (i.e. the height dimension is flattened) and allows an analysis of image distance (ego-centric range), image width (relative- in degrees, or absolute- in metres) and image direction (azimuth). With the new computer GUI, the graphical description applies to both the source images (where the recorded instrument seems to exist in the sound scene) and the reverberance images (the imagery created by reverberation in the recording environment). This allows a quantitative analysis of how the source and reverberance imagery is affected by the upmixing process.

5.1.1. Method and stimuli

A 30 second excerpt of an anechoicly recorded sung voice and (separately recorded) solo viola was reproduced with a single loudspeaker in a 2000 m^3 concert hall (RT60$\approx$ 3 s) and recorded with a pair of cardioid microphones, 16 cm spaced and angled at 110° (i.e. the ORTF configuration), 3.5 m from the source. The loudspeaker was either 1 m off the central axis (i.e. the axis equidistant to each microphone) or on-axis; as shown in table 1.

Fragment #	Source	Loudspeaker position
1	Viola	3 m left
2	Viola	3 m left
3	Voice	Centre
4	Voice	Centre

Table 1: Stimuli used in image-mapping and preference experiment. These recordings are the same as used in the image mapping experiment; made using an ORTF arranged mike pair in Pollack hall. The sound source was a loudspeaker on stage reproducing an anechoicly recorded solo music performance. The loudspeaker position was either equidistant to each microphone (i.e. "centre") or 3 m off-centre. Spectral and temporal details of the stimuli are given in [34].

For the upmixed scene, the setup was as shown in figure 5: in addition to the conventional front loudspeaker pair at ±30°, two rear loudspeakers were used at ±120° (and a delay was added to the front loudspeaker signals to account for the processing latency of the upmix system).[4] The loudspeakers were occluded using a visually opaque yet acoustically transparent curtain, 1 m from the central listening position, with markers at 10° intervals. The listener used the computer-driven GUI to map the locations of perceived source and reverberance sound images.

11 subjects took part in the experiment; all were sound recording students and were familiar using the mapping tool. Responses for each sound scene were overlayed to create *density plots*- as shown in table 2. Image directional strength (IDS) plots were created from these density plots, which show the number of times a source or reverberance image was elicited at a certain direction- as shown in table 3.

For each sound scene presentation the following source and reverberance image spatial properties were computationally extracted from the elicited image mappings:

- Source image azimuth.
- Reverberance image azimuth.
- Source image width (in degrees).
- Reverberance image width.
- Source image distance.
- Reverberance image distance.

Whether these image attributes were affected by the upmixing process was investigated using an ANOVA, and the results are shown in table 4.

5.1.3. Discussion

Looking at the density plots in table 2, it is difficult to identify any discernable trends in the source imagery between the upmixed 2/2 scene and original 2/0 scene. There is a similar range in reported image **distance**, and the ANOVA confirms that the difference in distance is not significant. As can be seen in the image direction strength plots in table 3, the source image **width** was unaffected by

[4] Bang & Olufsen Beolab 4000, two-way self-powered loudspeakers were used.

5.1.2. Results of image mapping experiment

Density plots:

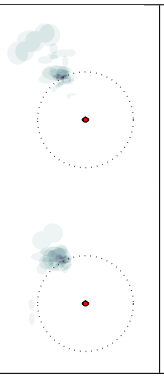
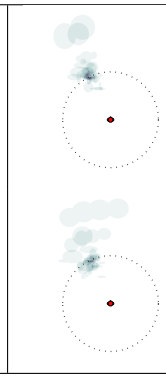
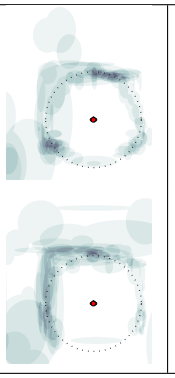
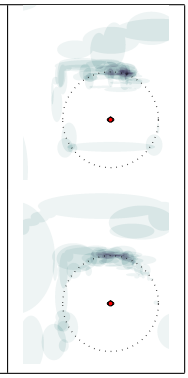
Image type	Stimulus	Scene config. 2/2	Scene config. 2/0
Source	Voice		
	Viola		
Reverb.	Voice		
	Viola		

Table 2: Density plots showing elicited source images and reverberance images in the upmixed 2/2 scene and reference 2/0 scene. Each of the unique audio scenes was graphically described by 11 subjects 3 times- i.e. there are 33 scene descriptions for each density plot. Original recording made using ORTF pair with anechoic voice or violin reproduced through a loudspeaker in a concert hall.

Image directional strength:

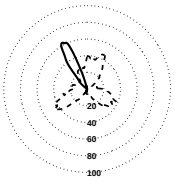
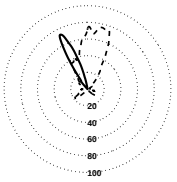
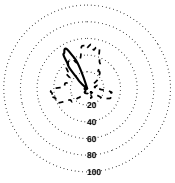
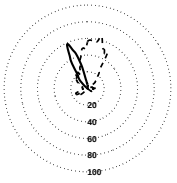
Stimulus	2/2	2/0
Voice		
Viola		

Table 3: Image Directional Strength: Showing the percentage of instances a listener reported stable source image (solid line) or reverberance image (dashed line) coming from a particular direction. These plots ignore the reported image distance and depth. Data has been smoothed with a Hanning-shaped window of width 5° (i.e. ±2.5°) to account for the spatial resolution of the auditory system for sound localization in the horizontal plane, which has a localization accuracy for broadband sources reproduced with loudspeakers around the listener of 3°-10° [26, pg. 41]. The radial scale corresponds to the percentage of times an image was reported in a particular direction from all 11 (subjects)×3 (runs) graphical descriptions of each unique audio scene (the outer-most circle on each polar plot corresponds to 100%).

Dependant variable:	Voice	Viola
Source image azimuth	$p = 0.000$	$p = 0.032$
Reverb. image azimuth	$p = 0.027$	$p = 0.004$
Source image width	$p = 0.243$	$p = 0.164$
Reverb. image width	$p = 0.000$	$p = 0.001$
Source image distance	$p = 0.918$	$p = 0.646$
Reverb. image distance	$p = 0.603$	$p = 0.337$

Table 4: Results of ANOVA analysis to see statistical significance of affect of scene configuration (2/0 or 2/2) on changes in S and R image spatial properties.

the upmixing process; with a mean image width of approximately 10°. The source image **azimuth**, however, was deemed significantly affected by the upmixing process. This trend is subtle but can be seen from the density plots; the source images in the upmixed scene were further to the left. This may be due to "leakage" of source-image components being radiated from the rear loudspeakers, distorting the source image (a *detent* effect). This can occur if the upmixer does not respond fast enough to transients, which can be mitigated by introducing a delay on the input signal relative to the analysis system; as used in the Fosgate system [21] (such a "look-ahead" analysis approach was not used in the experiments with the new upmix system). However, it is very promising to note that the source image was never reported in the direction of the rear loudspeakers in the upmixed sound scene; suggesting that the new upmix system can effectively remove sound components which affect source imagery from the rear loudspeaker signals.

The increased distribution (homogeneity) of reverberance imagery around the listener can clearly be seen for the upmixed scene. However, especially for the voice excerpt the R imagery is generally localized in the direction of the rear-left loudspeaker. This detent effect may be due to the complex transient nature of the sound, as transient sounds are easier to locate than sustained sounds [39, pg. 241].

5.2. Preference experiment

5.2.1. Method and stimuli

This experiment was conducted at the Banff Centre in an acoustically treated editing and mixing room (approximately 50 m^3, estimated RT_{60}<0.5 s). As with the previous experiment, four loudspeakers[5] were used, arranged in the conventional 2/2 ITU-775 [40] configuration (no centre-speaker), with rear loudspeakers at ±120°. The listener sat on a non-rotating chair at the sweet-spot, 2.3 m from each loudspeaker. The loudspeakers were calibrated so as to produce an equal SPL at the listening position (74±0.5 dB, unweighted, slow time averaging, using pink noise). The stimuli were burnt onto a DVD disc (recorded at 44.1 kHz, 16 bit), and the music was presented with four loudspeakers from a DVD-A player, as shown in figure 3.

The method of paired comparison was used to evaluate the new upmixer in terms of overall preference

[5]Type 1031 manufactured by Genelec.

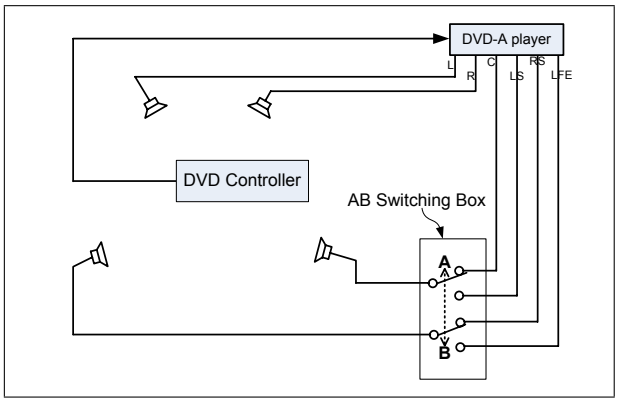


Fig. 3: Signal schematic for preference experiment. All audio outputs are from a DVD-A player (the LFE channel is a normal, full-bandwidth signal). The rear-loudspeaker channel switching device is a passive switching box with an "A-B" clicking switch. The channel gain-trims are not shown.

(as recommended by IEC-268 [41]). For each trial, the subject was presented two stimuli, which the subject could freely switch between using a 4-in, 2-out audio signal switching box, and to report which sound scene was preferred. Stimulus A or B corresponded to one of four scenes; a variant of the new upmixer or the 2/0 scene. This AB preference method was the same as used in the evaluation of various upmix systems in two other studies [22, 32].

Two groups of people undertook the experiment: 5 audio engineers and 11 musicians. The engineers were all past Tonmeister students, each with at least three years of experience with sound recording practice. The musicians were enroled on an intensive music performance or composition program at the Banff Centre (most of whom are professional).

Stimuli:
The original stimuli were the same as used in the previous GUI experiment. The difference in this preference experiment was that there were three variants of the new upmix system which were compared with each other and with the reference 2/0 scene. These three variants, plus the 2/0 scene, are now summarized:

1. Unmodified 2/2 upmixer (as used in the previous GUI experiment).
2. 2/0 (only the front two loudspeakers are 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 subject could advance to the next DVD track at any time by pressing the "next track" button, or could replay the track by pressing the "repeat track" button (i.e. the subject could take as long as they want to listen to the tracks and decide whether A or B was preferred)- as summarized in figure 3. The DVD play-mode was random, so the trial order was randomized (the subject would write down the track playing order on a piece of paper). Each of the four fragments were presented with each of the four scene configurations. Therefore there were 6 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.

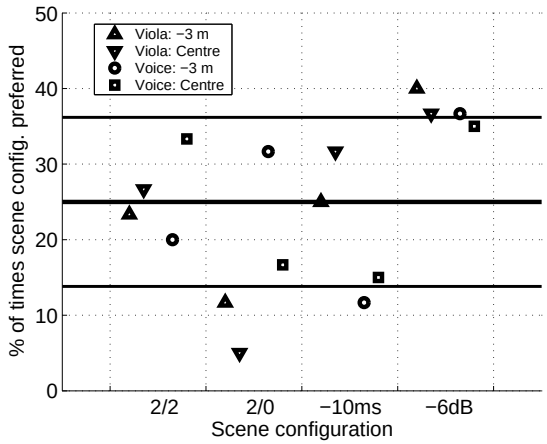
Subject task:
The subject was presented the twenty-four excerpts of music twice, with a 5-20 minute break in between. Using a computer GUI, they were asked: "`Which sound scene do you prefer: A or B?`" Once they selected either option, a pop-up window prompted the subject to confirm and advance the DVD to the next track. As mentioned, it was emphasised that the audio system they were evaluating was intended for use in a domestic home environment for entertainment purposes, so they should think about the preference task as if they were evaluating a product which they might purchase for home entertainment.

5.2.2. Results of preference experiment

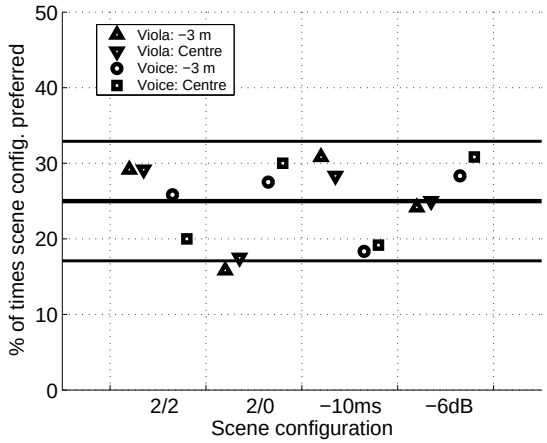
Results for the paired comparisons are shown in figure 4; which shows how often a particular stimulus was (dis)preferred over an other. The data is split into the musician and engineer group. The total number of trials (nTrials) was 24 scene pairs×2 runs×(the number of subjects), which was 240 for the engineer group and 480 for the musician group. 95% confidence intervals ($\pm 2\sigma$) were calculated according to (3):

$$\sigma = \sqrt{\text{nTrials} \times P(A)(1-P(A))}, \qquad (3)$$

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



(a) Audio engineer group (5 subjects).



(b) Musician group (10 subjects).

Fig. 4: Plot of percentage of preferred choices out of 4 possible scene configurations. 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). 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% CI's. If the marker is above the upper-most line, then this scene configuration was preferred significantly more than the others. 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 4 stimuli and 2 presentations; so 240 responses for the audio engineer group and 480 for the musician group.

5.2.3. Discussion

From the results of the preference choice analyses shown in figure 4, it can be seen that for the musician group there is only one statistically significant trend: for the viola recording at -3 m, which for the 2/0 scene was reported as being *less* preferred than the other scenes; this stimulus was also generally *dispreferred* by the engineer group. This lack of clear preference for the musician group was surprising, considering the large difference in sound scenes (such as the 2/0 and 2/2 comparisons). However, even though the audio-engineer group preferred the upmixed scene significantly more than the 2/0 scene (except for the voice at -3 m), which is a principle subjective design criteria, the result should be not be interpreted too generally as the engineer group are simply not "average listeners" in terms of experience. This finding is different from a study by Rumsey [32], who found that from a listening panel of 22 experienced listeners (Tonmeister students and professional audio engineers), a conventional 2/0 audio scene was generally preferred *more* than an upmixed 3/2 scene for music recordings, though there was no interpretation of the statistical significance of the preference choices in his study.

It is particularly interesting that the engineer group generally preferred the 2/2 scene with the 10 ms delay *less* than the other scenes; significantly less for the -3 m viola 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 R image collapsing to the front speakers due to the precedence effect. Looking at the cross-correlation analysis of the side loudspeaker channels $R-RS$ (figure 2(b)), it can be seen that these channels have a non-zero correlation. As figure 4 shows, the scene with the -6 dB rear channels was significantly preferred over all other scenes for the engineer group, though this was close to the level of chance for the centre-voice stimuli.

6. CONCLUSION

A new audio upmixing system to enhance reverberance (or "ambiance") imagery for the reproduction of conventional unencoded two-channel music recordings was proposed and evaluated. The new upmixer uses a novel arrangement of adaptive filters and signal delays, and was evaluated according to three criteria relating to perceived source and reverberance imagery, and overall preference (the upmixed 2/2 system was compared with the original 2/0 scene). Electroacoustic measurements of the output signal correlation were also undertaken to provide a theoretical validation of theory describing the system behavior for different input signals.

The filter system was able to align input signals in terms of both frequency (typically, with a 1024-point resolution) and time (typically, with a ± 10 ms tolerance). This is in contrast with all extant upmix systems, whose performance is reduced when spaced microphone or time-delay panned recordings are used as these upmixers do not have a mechanism for time-aligning the input signals before the difference signal is calculated.

It was shown that due to the self-orthogonalizing properties of the adaptive filter (updated according to the normalized-least-means-square algorithm), diagonally opposite loudspeaker signals in the upmixed scene had a near-zero correlation, whilst side loudspeakers had a non-zero correlation. It is argued that this would maximize spatial fidelity of the source image in the upmixed scene, whilst creating reverberance images localized to the sides of the listener, and that the upmixed scene would be preferred over the original 2/0 scene. These three assertions were shown to be valid for a variety of two-microphone recordings of solo musical instrument performances in a concert hall. Furthermore, a study with experienced listeners showed that four-loudspeaker audio scenes created with the new upmix system were generally preferred over the original two-loudspeaker scene.

7. ACKNOWLEDGEMENTS

This work was sponsored by Bang and Olufsen and a grant from the National Sciences and Engineering Research Council of Canada. The experiments were undertaken at McGill University; the Banff Centre; and Philips Research, and many thanks to Theresa Leonard; Steve Bellamy; and Dr. Ronald Aarts for their advice and assistance with this. The PhD thesis from which this work comes [34] was undertaken at McGill University, and the author is indebted to the help received from his advisors: Professors W. Martens; W. Woszczyk; J. Benesty and A. Bregman.

References

[1] B. G. Shinn-Cunningham. The real reasons you should invest in a surround-sound system. *Journal of the Acoustical Society of America*, 119(5):3280, 2006.

[2] G. Theile and G. Plenge. Localization of lateral phantom sources. *Journal of the Audio Engineering Society*, 25(4):196–200, 1977.

[3] J. Usher. Design criteria for high quality upmixers. In *Proceedings of the AES 28th international conference*, Piteå, Sweden, 2006.

[4] F. Rumsey, Zieliński, R. Kassier, and S. Bech. On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *Journal of the Acoustical Society of America*, 118(2):968–976, 2005.

[5] N. Zacharov and K. Koivuniemi. Unravelling the perception of spatial sound reproduction: analysis and external preference mapping. In *Proceedings of the AES 111th international convention*, New York, 2001.

[6] M. Morimoto. How can auditory spatial impression be generated and controlled? In *Proceedings of the International Workshop on Spatial Media*, Aizu-Wakamatsu, Japan, 2001.

[7] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross. Objective measurements of listener envelopement in mulitchannel surround systems. *Journal of the Audio Engineering Society*, 51(9):826–840, 2003.

[8] J. Usher and W. Woszczyk. Design and testing of a graphical mapping tool for analyzing spatial audio scenes. In *Proceedings of the AES 24th International Conference on Multichannel Audio*, Banff, Canada, July 2003.

[9] J. Usher and W. Woszczyk. Visualizing auditory spatial imagery of multi-channel audio. In *Proceedings of the AES 116th international convention*, Berlin, Germany, 2004.

[10] J. Usher, W. L. Martens, and W. Woszczyk. The influence of the presence of multiple sources on auditory spatial imagery. In *Proceedings of the 18th International Congress on Acoustics*, Kyoto, Japan, April 2004.

[11] J. Usher and W. Woszczyk. Interaction of source and reverberance spatial imagery in multichannel loudspeaker audio. In *Proceedings of the AES 118th international convention*, Barcelona, Spain, 2005.

[12] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel. Sound source localization in a five-channel surround sound reproduction system. In *Proceedings of the AES 107th international convention*, New York, 1999.

[13] V. Pulkki and M. Karjalainen. Localization of amplitude-panned virtual sources I: Stereophonic panning. *Journal of the Audio Engineering Society*, 49:739–752, 2001.

[14] J. Corey and W. Woszczyk. Localization of lateral phantom images in a 5-channel system with and without simulated early reflections. In *Proceedings of the AES 113th international convention*, Los Angeles, 2002.

[15] C. Guastavino and B. Katz. Perceptual evaluation of multi-dimensional spatial audio reproduction. *Journal of the Acoustical Society of America*, 116(2):1105–1115, 2004.

[16] C. Avendano and J.-M. Jot. A frequency-domain approach to multichannel upmix. *Journal of the Audio Engineering Society*, 52(7/8):740–749, 2004.

[17] M. A. Gerzon. Optimum reproduction matrices for multispeaker stereo. *Journal of the Audio Engineering Society*, 40(7/8):571–589, 1992.

[18] M. T. Miles. An optimum linear-matrix stereo imaging system. In *Proceedings of the AES*

101st international convention, Los Angeles, 1996.

[19] P. Scheiber. Quadrasonic sound system. US patent # 3,632,886, 1972.

[20] D. Griesinger. Multichannel matrix surround decoders for two-eared listeners. In *Proceedings of the AES 101st international convention*, 1996.

[21] K. Gundry. A new active matrix decoder for surround sound. In *Proceedings of the AES 19th international conference*, Schloss Elmau, Germany, 2001.

[22] R. Irwan and R. M. Aarts. Two-to-five channel sound processing. *Journal of the Audio Engineering Society*, 50(11):914–926, 2002.

[23] R. Plomp and A. M. Mimpen. Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech-reception threshold for sentences. *Acustica*, 48:325–328, 1981.

[24] B. G. Shinn-Cunningham, J. Schickler, N. Kopco, and R. Y. Litovsky. Spatial unmasking of nearby speech sources in a simulated anechoic environment. *Journal of the Acoustical Society of America*, 110:118–1129, 2001.

[25] M. L. Hawley, R. Y. Litovsky, and J. F. Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America*, 115(2):833–843, 2004.

[26] J. Blauert. *Spatial hearing: The psychophysics of human sound localization*. MIT Press, Cambridge, Mass., 1997.

[27] M. Barron. The subjective effect of first reflections in concert halls - The need for lateral reflections. *Journal of Sound and Vibration*, 15:475–494, 1971.

[28] M. Morimoto. The relation between spatial impression and the precedence effect. In *Proceedings of the International Conference on Auditory Display*, 2002.

[29] K. Hiyama, S. Komiyama, and K. Hamasaki. The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field. In *Proceedings of the AES 113th international convention*, Los Angeles, 2002.

[30] D.G. Malham. Homogeneous and nonhomogeneous surround sound systems. In *Proceedings of the AES UK conference: Second Century of Audio*, London, 1999.

[31] J.-J. Sonke. *Variable acoustics by wave field synthesis*. PhD thesis, TU Delft, 2000.

[32] F. Rumsey. Controlled subjective assessments of two-to-five-channel surround sound processing algorithms. *Journal of the Audio Engineering Society*, 47(7/8), 1999.

[33] S. Makino. Blind source separation of convolutive mixtures of speech. In J. Benesty and Y. Huang, editors, *Adaptive signal processing*, chapter 7, pages 195–225. Springer, New York, 2003.

[34] J. S. Usher. *Subjective evaluation and electroacoustic theoretical validation of a new audio upmixer*. PhD thesis, McGill University, Schulich school of music, 2006.

[35] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, Englewood Cliffs, N. J., 4th edition, 2001.

[36] R. Mason, T. Brookes, and F. Rumsey. Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli. *Journal of the Acoustical Society of America*, 117(3):1337–1350, 2004.

[37] R. Mason, N. Ford, F. Rumsey, and B. de Bruyn. Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction. *Journal of the Audio Engineering Society*, 49(5):366–384, May 2001.

[38] N. Ford. *Developing a Graphical Language to Represent Listeners' Experiences of Spatial Attributes in Reproduced Sound*. PhD thesis, Univerity of Surrey, England. School of Performing Arts, 2005.

[39] B. C. J. Moore. *An introduction to the psychology of hearing*. Academic Press, San Diego, Calif., 4th edition, 1997.

[40] ITU-R BS. 775-1. Multichannel stereophonic sound system with and without accompanying picture. Recommendation BS.775-1, International Telecommunication Union Radiocommunication Assembly, 1992-1994 1994.

[41] IEC 268-13. Sound system equipment–part 13: Listening tests on loudspeakers. Technical report, International Electrotechnical Commission, Geneva, Switzerland, 1985.

[42] J.-M. Jot and A. Chaigne. Analysis and synthesis of room reverberation based on a statistical time-frequency model. In *Proceedings of the AES 103rd international convention*, New York, 1997.

[43] B. Blesser. Interdisciplinary synthesis of reverberation viewpoint. *Journal of the Audio Engineering Society*, 49(10):867–903, 2001.

[44] B. Widrow and M. E. Hoff. Adaptive switching circuits. In *IRE WESCON Convention Record*, pages 96–104, 1960.

[45] S. Haykin and B. Widrow. *Least-Mean-Square Adaptive Filters*. Wiley, New-York, 2003.

[46] S.L. Gay. The fast affine projection algorithm. In S.L. Gay and J. Benesty, editors, *Acoustic Signal Processing for Telecommunication*, chapter 2. Kluwer Academic Publishers, Boston, 2000.

[47] 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. on Circuits and systems*, 34(7):788–798, 1987.

[48] B. Widrow and J. M. McCool. Stationary and nonstationary learning characteristics of the LMS adaptive filter. In *Proceedings of the IEEE*, volume 64, pages 1151–1162, 1976.

[49] D. T .M. Slock. On the convergence behaviour of the LMS and the normalized LMS algorithms. *IEEE Trans. on Signal Processing*, 41(9):2811–2825, 1993.

APPENDIX

1. IMAGERY CREATED BY THE NEW UPMIXER

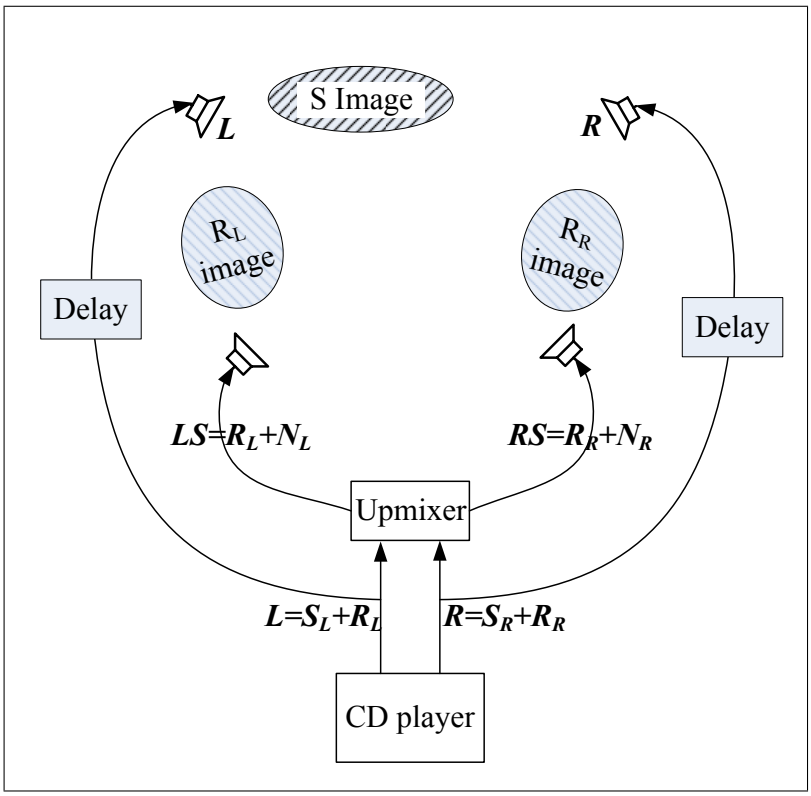


Fig. 5: In this idealized summary it is shown that the system processes two input signals (L and R) of a musical recording (such as the output of a CD player) and creates two new signals LS and RS which are then radiated by the rear loudspeakers. The front loudspeakers radiate the original signals with a time delay to account for the IO latency of the DSP system. It is assumed that those components in the original signals which are correlated contribute to the formation of the source image; these components are S_L and S_R in the left and right channel from the CD player. The remaining components in each channel contribute to the perceived reverberance image (these sound components are R_L and R_R). The upmixer extracts these reverberance image components and radiate them from the rear speakers, creating reverberance virtual (phantom) images to the side of the listener. Other "noise" artifacts created by the new upmixer N_L and N_R, are also radiated. To avoid distortion of the source image there should be no source image components in the rear loudspeaker channels; that is, signal RS should be uncorrelated with signal L, and LS uncorrelated with R.

2. OVERVIEW OF THE NEW UPMIXER

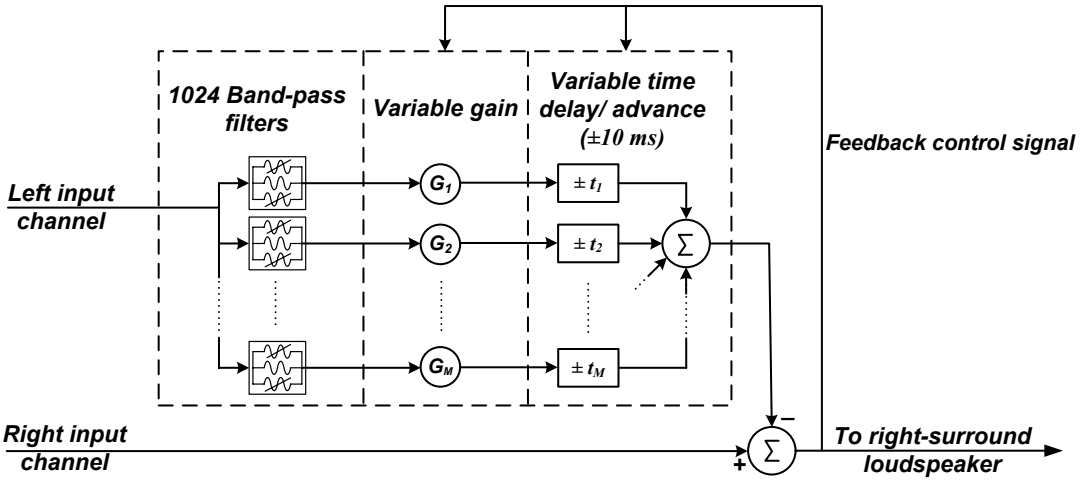


Fig. 6: Conceptual overview of system, for generation of one surround channel. M equal band-width band-pass filters (typically, M=1024) are weighted by a positive or negative constant and the filtered channel can then be time advanced or delayed to minimize the difference signal (the differencing is actually done on a frequency-by-frequency basis). In the actual implementation, the right-input signal is delayed by about 10 ms, and the variable delay operates between a delay of 0 and about 20 ms.

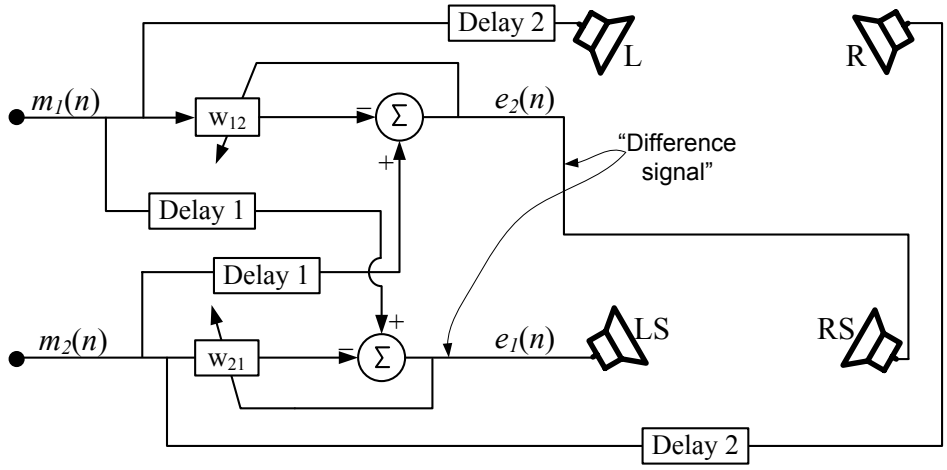


Fig. 7: Electroacoustic signal chain of two the new upmixer, showing the signals feeding the four loudspeakers (which are intended to be arranged according to the ITU-775 2/2 [40] recommendation). Delay 1 and delay 2 are different: the former is to allow for non-minimum phase impulse responses when the direct sound arrives in one channel before the other, and the latter is to account for IO delay caused by the DSP system. Not shown in this diagram is the cross-talk on the input signals which improves performance when the input audio contains hard-panned sources, as explained in section 5.4 of the dissertation [34].

3. UPDATE OF THE ADAPTIVE FILTER (NLMS ALGORITHM)

With reference to the signal schematic in figure 7, we see that each input signal m_1 and m_2 is filtered by an M-sample length filter ($\mathbf{w}_{21}$ and $\mathbf{w}_{12}$, respectively). These filters model the early component of the impulse response between the two microphone signals; so ideally M is equal to the mixing time (i.e. the time index of the room impulse response which can be approximates a random ergodic process like decaying white noise [42, 43], which is taken as 10 ms). A delay is added to input channel m_i before the filtered signal y_i is subtracted; this is to allow for non-minimum phase impulse responses which can occur if the sound source is closer to one microphone than the other. However, for the foregoing analysis we will not consider this delay as it makes the mathematical description more straightforward (and it would make no difference to the theory if it was included).

The filtering of signal m_j by the adaptive filter $\mathbf{w}_{ij}$ gives signal $y_i(n)$. This subscript notation may seem confusing, but helps describing the loudspeaker output signals because signal m_i and e_i are both phase-coherent (have a non-zero correlation) and are reproduced by loudspeakers on the same side (e.g. signals m_1 and e_1 are both reproduced with loudspeakers on the left-hand side). This filtering process is summarized as the discrete time linear convolution in (4):

$$y_i(n) = \sum_{k=0}^{M-1} m_j(n-k) w_{ij,k}, \tag{4}$$

which with the following definitions:

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

allow the convolution to be written in vector form as:

$$y_i(n) = \mathbf{m}_j^T(n)\mathbf{w}_{ij}. \tag{5}$$

If we look at filter $\mathbf{w}_{12}$ in figure 7, it is seen that the filtered m_2 signal, y_1 is subtracted from the unfiltered m_1 signal (sample-by-sample) to give the error signal e_1:

$$e_i(n) = m_i(n) - y_i(n). \tag{6}$$

The output signal is conventionally called an error signal [44, 45] as it can be interpreted as being a mismatch between y_i and m_i caused by the filter coefficients $\mathbf{w}_{ij}$ being "not-good enough" to model m_i as a linear transformation of m_j; these terms are used for the sake of convention and these two error signals are the output signals of the system which are reproduced with separate loudspeakers behind the listener.

If the filter coefficients $\mathbf{w}_{ij}$ can be adapted so as to approximate the early part of the inter-microphone impulse response, then the early-correlated sound component will be removed and the "left-over" signal will be the reverberant (or reverberance-image) component in the m_j channel, plus a filtered version of the reverberant component in the m_i channel. In this case, the error signal level will be smaller than the original level of m_j. The "goal" of the algorithm which changes the adaptive filter coefficients can therefore be interpreted as to minimize the level of the error signals. This level can simply be calculated as a power estimate of the output signal $\mathbf{e}_i$, which is an average of the squares of the individual samples, and it is for this reason that the algorithm is called the Least Mean Square (LMS) *algorithm* [44, 45]. This goal is formally expressed as a "performance index" or "cost" scaler J, where for a given filter vector $\mathbf{w}_{ij}$:

$$J_i(\mathbf{w}_{ij}) = E\left\{e_i^2(n)\right\}, \tag{7}$$

and $E\{\cdot\}$ 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" [35].

When a filter is in the optimal state, the rate of change in the error signal level (i.e. J) with respect to the filter coefficients $\mathbf{w}$ will be minimal. This rate of change (or gradient operator) is a M-length vector ∇, and applying it to the cost function J gives:

$$\nabla J_i(\mathbf{w}_{ij}) = \frac{\partial J_i(\mathbf{w}_{ij})}{\partial \mathbf{w}_{ij}(n)}. \tag{8}$$

The right-hand-side of (8) is expanded using partial derivatives in terms of the error signal $e(n)$:

$$\frac{\partial J_i(\mathbf{w}_{ij})}{\partial \mathbf{w}_{ij}(n)} = 2E\left\{\frac{\partial e_i(n)}{\partial \mathbf{w}_{ij}(n)} e_i(n)\right\}, \tag{9}$$

and the general solution to this differential equation, for any filter state, can be obtained by first

substituting (5) into (6):

$$e_i(n) = m_i(n) - \mathbf{m}_j^T(n)\mathbf{w}_{ij}(n) \tag{10}$$

and then differentiating with respect to $\mathbf{w}_{ij}(n)$:

$$\frac{\partial e_i(n)}{\partial \mathbf{w}_{ij}(n)} = -\mathbf{m}_j(n). \tag{11}$$

This gives us the differential expression on the right-hand side of (9) and allows the filter update term in (8) to be solved as:

$$\nabla J_i(\mathbf{w}_{ij}) = -2E\left\{\mathbf{m}_j(n)e_i(n)\right\}. \tag{12}$$

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. The expectation operator in equation (12) is replaced with a vector multiplication and the filter update (or the steepest descent gradient algorithm) is:

$$\mathbf{w}_{ij}(n) = \mathbf{w}_{ij}(n-1) + \mu\mathbf{m}_j(n)e_i(n). \tag{13}$$

It should be noted that the adaptive filtering algorithm which is used (i.e. based on the LMS algorithm) is chosen because of its relative mathematical simplicity compared with others (such as the affine projection; 46 or RLS; 35 algorithms), yet it is powerful enough to satisfy both the subjective and electronic design criteria.

Besides the massive increase in computational efficiency of implementing the filter-update and signal filtering in the frequency domain (requiring 5 FFT's per iteration; i.e. for every M input samples), the performance of the frequency-domain and time-domain NLMS algorithm are equivalent [47]. The overlap-save technique was used (as described in [47]) with an overlap factor of two (performance was not significantly affected by an increase in overlap).

From the filter update equation (13) 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 gradient ∇ in (8) should be zero but the actual ∇ 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* [48] which is proportional to the level of $\mathbf{m}_j(n)$. This undesirable consequence can be mitigated by normalizing the gradient estimation with another scaler which is inversely proportional to the power of $\mathbf{m}_j(n)$, and the algorithm is therefore called the Normalized Least–Mean–Square (NLMS) algorithm [49]. The tap-weight adaptation is then:

$$\mathbf{w}_{ij}(n) = \mathbf{w}_{ij}(n-1) + \frac{\alpha}{\delta + \mathbf{m}_j^T(n)\mathbf{m}_j(n)}\mathbf{m}_j(n)e_i(n), \quad \text{with} \quad 0 < \alpha < 1. \tag{14}$$

When the input signals $\mathbf{m}_1(n)$ and $\mathbf{m}_2(n)$ are very small, inverting the power estimate could become computationally problematic. Therefore a small constant δ is added to the power estimate in the denominator of the gradient estimate- a process called *regularization* [35, pg. 338]. How the regularization parameter affects filter convergence properties was investigated empirically using a variety of music audio signals [34].

3.1. The Principle of Orthogonality

As mentioned, when the "optimal state" is attained when the gradient operator is equal to zero, so under these conditions at sample time n, (12) becomes:

$$E\left\{\mathbf{m}_j(n)e_i(n)\right\} = \mathbf{0}_{M\times 1}. \tag{15}$$

This last statement represents the Principle of Orthogonality (PoO) [35, pg. 96]. The elegant relationship means that when the optimal filter state is attained, referring back to figure 7, e_1 (the rear-left loudspeaker signal) is uncorrelated with m_2 (the front-right loudspeaker signal). This means that when the adaptive filter is in its optimal solution, diagonally opposite loudspeaker signals are uncorrelated: *Quod Erat Demonstrandum.*

Under such a optimal operating conditions (i.e. when the error signal level is minimal- also called the *Wiener solution*), distortion of the source image by the upmixer is minimized because signal e_i contains reverberance-image components which are unique to m_i, and as the source image is only affected by correlated components within m_i and m_j (by definition; correlated components within an approximately 20 ms window), then a radiated signal which is uncorrelated with *either* m_i or m_j can not contain a sound component which affects source imagery.