Design criteria for high quality upmixers

John Usher

Centre for Interdisciplinary Research in Music Media and Technology, McGill University, Montreal, Canada

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

ABSTRACT

Audio signal processing systems for converting two-channel (stereo) recordings to four or five channels are increasingly relevant. These blind upmixers can be used to reproduce conventional stereo recordings with multichannel home-theatre or automotive spatial audio systems to create a more engaging and natural listening experience in terms of auditory spatial imagery. This paper discusses a number of design criteria for ensuring the upmixed sound scene maintains a spatial image fidelity true to the original 2/0 audio scene. These subjective criteria are presented as a method which can be analysed using both listening tests and electroacoustic measurements. Electronic performance of three upmixers are compared with respect to these criteria using stimuli containing both time-delay and amplitude panning.

1. INTRODUCTION

The quality of reproduced sound using loudspeakers has been increasing at a steady rate for over a century. In terms of timbral recreation of a recorded sound, there is a strong argument for saying it is as good as it is going to get. However, "spatial quality... has some way to go before the curve could be said to be asymptotic to some ideal" [1]. This discrepancy is due to the relatively new arrival of multichannel audio systems in our homes and cars, providing the means to reproduce sound in a way which is more engaging and aesthetically "natural". And yet the vast majority of our musical recordings are stored on a two-channel "stereo" format which we are forced to listen to with a two-loudspeaker electroacoustic system. This paper discusses a class of audio signal processors which enables reproduction of these two-channel recordings with surrounding loudspeakers. Such an *upmixing* device can be classed as a spatial audio enhancer.

Idiomatically, a spatial audio system means a sound reproduction system which is designed to *enhance* a listening experience in terms of spatial sound quality. Enhancement of a listening experience may be measured according to a number of criteria, for example an enhanced speech intelligibility or perceived naturalness. In the present work, *enhancement* is related to changes of spatial imagery in a listening experience created using a conventional loudspeaker pair (a 2/0 system; [2]) in ways which a listener can describe in terms of perceived spatial imagery.

Of course, audio signal processing will generally affect auditory imagery in terms of *timbre* as well as space (indeed, auditory imagery is often taken as meaning a representation of perceived sound in terms of these two aspects [3, 4]). However, there is strong evidence that the spatial sound quality significantly affects overall quality ratings [5]. For concision, the upmixer design criteria presented here are limited to the spatial aspects of imagery.

The assumption maintained in this paper is that the underlying imperative governing an upmixer design is that the sound imagery evoked be consistent with that in a conventional two-loudspeaker sound scene created using the same recording. As will be shown, this general imperative is translated into meaning that the spatial imagery associated with the recorded musical instrument (the *source* image [6]) remains the same (undistorted) in the upmixed sound scene. The enhancement is therefore in terms of the imagery which contributes to the listeners' sense of the recording space; the *environment*, *ambiance* or *reverberance* imagery [6, 7].

The enhancement can be qualified in two ways: firstly, using electroacoustic measurements which relate to parameters known to affect imagery in audio (such as signal cross-correlation). And secondly, with subjective listening tests; for example using a descriptive comparison between the imagery of the upmixed sound scenes and a conventional two-loudspeaker audio scene or a preference experiment where sound scenes created with various configurations of the new upmixer are compared with a conventional 2/0 audio scene.

The design aim of a high quality upmixer can be summarised with three general goals relating to spatial imagery. These goals relate to modifying the listening experience of a conventional loudspeaker-pair reproduction of a musical recording:

1. To create a source image with a spatial quality similar to the original 2/0 mix.
2. So as to create a natural-sounding ambiance (reverberance) image.
3. To create a listening experience which people would prefer over the original 2/0 listening experience.

The third goal is assumed subservient to the first two; such high quality upmixers are not intended as a "special effect" which reinterprets the mixing intention of the recording producer, but rather as a system to compliment these intentions in ways which are consistent with sound in the natural environment.

2. SUBJECTIVE DESIGN CRITERIA

The design criteria which can be evaluated with listening tests are outlined here. As mentioned in the introduction, the upmix system should enhance the perceived spatial imagery of a conventional two-channel recording reproduced with a 2/0 loudspeaker pair. It should be bared in mind that the upmix system is considered solely as a signal processing-based solution, not a new transduction mechanism. This ensures compatibility with existing home (or automotive) theatre systems, such as those arranged according to the 2/2 ITU-R BS.775-1. recommendation. The 2/0 reproduction is therefore the reference case which applies to the three criteria outlined below.

1. *Spatial distortion of the source image in the upmixed scene should be minimized (compared with 2/0 loudspeaker audition)* .
   Image attributes of the source image should be the same as with a 2/0 reproduction. How this can be quantitatively measured is summarised in the next section (e.g. using descriptive analysis methods). Another way of interpreting this goal is that the upmix system should be designed in such a way that it *respects the mixing intentions of the sound engineers involved in the production of the original two-channel recording*, at least in terms of the source ("front-stage") imagery. This criterium is also identified by Gerzon, in the context of non-surrounding upmixers for reproducing n_1 original channels with a larger number (n_2) of loudspeakers: "*the localization properties of the reproduced sound over n_2 loudspeakers should be as similar as possible to that originally intended via n_1 loudspeakers*" [8].
2. *Reverberance imagery should have a homogenous distribution in the horizontal plane; in particular, reverberance image strength should be high from lateral ($\pm 90^\circ$) directions.*
   It was found in a previous experiment [6] that when a reverberance image was panned at 90° using the front right and rear-right loudspeakers, the reverberance (R) images were reported as being spread out between these side speakers. Therefore, phantom imaging to the side can increase the perceived lateral reverberance image content of the sound scene, and lateral reflections have been shown to contribute to increased spatial impression [9]. Furthermore, when the R image was panned at 90° the source (S) image distortion was less than when it was panned at 30° or 60°. In other words, by re-panning the reverberance image to the side of a listener the frontal source images can be spatially unmasked, as measured by a reduction in spatial fusion between the S and R images [6].
3. *The upmix system should be preferred to a conventional 2/0 system.*
   In the context of a home musical listening experience (i.e. listening for *pleasure*)- the system should be preferred over a reference 2/0 reproduction created using the same recording. *Preference* is chosen rather than the similar (and for most cases, probably identical) constructs *naturalness* or *sound quality* because it is intended that the spatial enhancement provided by the system be apparent for the general population and preference is a less esoteric concept than naturalness or sound quality to use in listening tests with non-experienced ("naïve") listeners.

3. EVALUATION OF IMAGERY WITH LISTENING TESTS

There are a number of ways for evaluating the degree to which an upmixer meets the subjective design criteria previously outlined. Descriptive analysis (DA) [10, 11] approaches are often used. Here, sound *character* descriptions (i.e. judgements of sound attributes which are not affect by *sentimental* preferences [3]) are directly rated. This provides data which are easy to statistically interpret and allows a quantitative insight into differences between sound scenes, such as how the spatial properties of the source or reverberance imagery differs between an unprocessed 2/0 scene and the upmixed sound scene. DA is a verbal description technique, where the sound attributes are provided by the experimenter (the attribute list may be the result of previous experiments provided by the listeners themselves). Such verbal techniques have been criticized for their level of abstraction from the task in hand [12]. For instance, reporting the absolute image width of two sound creating objects is not as intuitive as pointing to the region of space the sound source seems to occupy. Furthermore, DA techniques suffer from cognitive bias effects which occur due to our experience with the recorded sounds in the real-world leading the listener to certain expectations about the spatial image. For instance, recordings of shouted speech are consistently reported farther than speech at a whispered or conversational level, even when the shouted speech is presented at a level over 6 dB greater [13].

Graphical mapping of imagery in reproduced sound scenes provides a direct way to describe the spatial envelope of auditory images [12, 14, 15]. 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). This 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 both the source and reverberance imagery is affected by the upmixing process.

3.1. Method and stimuli

As an example of a method for investigating how well the subjective design criteria are met, an experiment is now described which used a computer-driven graphical mapping system (a GUI) [6, 14] to compare a 2/0 loudspeaker scene with that created by a new 2-to-4 channel upmixer (details of this upmixer are given in [4]). The version of the upmixer used here was an early incarnation and the results presented here are not intended as a definitive evaluation of the system, but, in keeping with the theme of this paper, as a demonstration of how spatial auditory imagery in an upmixed scene can be compared to that in the original 2/0 scene. The new upmixer creates two new "ambiance" channels which are reproduced behind the listener as shown in figure 1; therefore, because the front loudspeaker channels are unaffected this system might be better described as an *ambiance extraction* device [16]. An electronic analysis of the output signals is compared with two commercial systems later in this paper. The new upmixer is given the acronym ASUS; Adaptive Sound Upmix System.

A 30 second excerpt of an anechoicly recorded sung voice and (separately recorded) viola was reproduced with a single loudspeaker in a 1600 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 1 m off the central axis (i.e. the axis equidistant to the each microphone). For the original reference sound scene, the two microphone channels were reproduced with a loudspeaker pair in front of the listener (i.e. the conventional 2/0 arrangement). For the upmixed scene, the setup was as shown in figure 1: in addition to the conventional loudspeaker pair, 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). 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 1. Also, 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 was affected by the upmixing process was investigated using an ANOVA, and the results are shown in table 2.

3.3. Discussion

Looking at the density plots in table 1, 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** (which is larger for the voice than viola), and the ANOVA confirms that the difference in distance is not significant. Likewise, the source image **width** was unaffected by 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 [17] (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.

Regarding the spatial distribution of the reverberance (R) images; it was surprising to find that on

3.2. Results

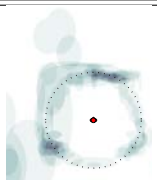
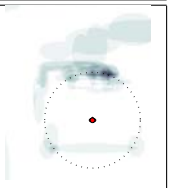
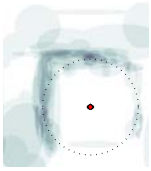
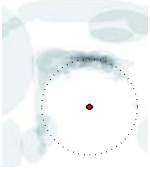
Image type	Stimulus	Scene config. 2/2	Scene config. 2/0
Source	Voice		
	Viola		
Reverb.	Voice		
	Viola		

Table 1: 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.

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 2: 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.

four occasions listeners reported R images *behind* them in the 2/0 scene (out of 66 different presentations). This may have been due to an error with the playback system (the GUI and playback system were on separate computers) or simply an auditory illusion. 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 located 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 [18, pg. 241].

In a further experiment [4], listeners preferred an upmixed scene with the rear channels attenuated by 6 dB to the version used in this experiment. It was interesting to note that even with experienced listeners, a delay of the rear-loudspeaker signals by 10 ms did not affect preference (such a delay is often used in upmixers to reduce fusion of source-image components present in the front and rear loudspeaker signals; e.g. [16, 19, 20]).

4. ELECTRONIC DESIGN CRITERIA

The subjective design criteria are now translated into a set of criteria which can be evaluated using electronic measurements. The criteria can be divided into two categories; those which concern source imagery and those which concern reverberance imagery. To describe how to realize these goals in signal processing terms the upmixer input signal is also modeled as two parts; a part which affects spatial aspects of the source imagery and a part that affects spatial aspects of reverberance imagery. How these two parts are distinguished in electronic (or acoustic) terms is discussed in [4, 21], but for now these two electronic components of the input signals are simply called the Source (S) component and the Reverberance (R) component. In the left input channel, these components are abbreviated to S_L and R_L, and in the right channel S_R and R_R. Other sound components which do not contribute to S or R imagery, i.e. noise in the recording environment from a source other than the musical instrument, are assumed to be absent or at least very low in level. Therefore the left and right input signals (i.e. the left and right channels from the recorded music, such as from a CD player) can simply be modeled as the sum of these two sound components- as summarised in figure 1. Although the analysis is only discussed for a solo instrument recording, the theory is generalizable to any number of recorded sources.

According to the principles of pair-wise panning [22], if the source components S_L and S_R are coherent (i.e. with a high absolute cross-correlation peak at a lag less than about 1 ms) then radiation of these signals with two loudspeakers either in front (as with a conventional 2/0 loudspeaker system) or to the side of the listener will create a phantom source image between the loudspeakers [6, 23]. The same applies to the the radiation of the reverberance components (e.g. created from a channel of artificial reverberation) [6]; so if R_S could be extracted from the right channel and radiated from the rear-right loudspeaker, a listener would perceive a reverberance image on the right-hand side, as shown in figure 1. As we are dealing with a noise free (or at least, very low noise) recording environment, the reverberance image components can simply be defined by exclusion: they are those sound components of the two input signals which are not correlated.[1]

Implicit in this upmix approach is that if there are *no* reverberance image components (e.g. for anechoic recordings, or with electronic sound sources), then there will be no ambiance to extract. On the other hand, when R image components *are* present, the overall sound level in the room will be greater in the upmixed scene; which violates the "energy preservation" criteria [8].

The electronic criteria are now summarized:

1. *Spatial distortion of the source image in the upmixed scene should be minimized compared with 2/0 loudspeaker audition.*
   To maintain the auditory spatial imagery of a source image perceived with the 2/0 reproduction, source image components L_S and R_S should not be radiated from the rear loudspeakers in the upmixed sound scene. 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

[1] At least, not correlated for lag-times within the first 10-20 ms, as shall be described shortly.

time n, this is mathematically 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), as it is the early sound after the direct-path sound which contributes to spatial aspects of source imagery (such as timbral colouration) and the later part to reverberance imagery ("spatial impression") [9, 24, 25]. For lag times (k) greater than 20 ms or so, the two signals may be somewhat correlated at low frequencies- as explained below.

2. *Reverberance imagery should have a homogenous distribution in the horizontal plane; in particular, reverberance image directional strength should be high from lateral ($\pm 90°$) directions.*
The implication of this statement 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 would 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 was found in [6]. This is summarized 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.

Regarding the degree of correlation between the two rear channels (i.e. the "extracted ambiance" signals), the optimal relationship is not as straightforward as with the above two electronic design criteria. Although low-frequency interaural coherence is conducive for enveloping, close-sounding and wide auditory imagery [26], this does not necessarily mean the rear loudspeaker channels should be uncorrelated *de facto*. The correlation between two locations in a reverberant field is dependant on the distance between them and is frequency dependant [27]. For instance, at 100 Hz the measuring points in a reverberant field must by approximately 1.7 m apart to have a coherence of zero (assuming the Schroeder frequency of the hall is less than 100 Hz). Microphone-pair recordings in concert halls therefore rarely have total decorrelation at low-frequencies. Furthermore, for sound reproduced with a loudspeaker pair in normal echoic rooms, due to loudspeaker cross-talk, head diffraction and room reflections the interaural coherence at low frequencies is close to unity regardless of the interchannel coherence of the loudspeaker signals [4, 28].

5. ELECTRONIC COMPARISON OF THREE UPMIXERS

In this section a comparison is reported with the output signal properties of the new upmixer (ASUS) and two commercially available upmixers. Various two-channel input signals were used, all of which are un-encoded (i.e. have not been processed with any down-mix algorithm). The new upmix system is a two-to-four channel upmixer, whereas commercial upmix systems are nearly all two-to-five (or more) channel systems which utilize the centre channel of the ITU-R BS.775-1 [2] loudspeaker configuration.

5.1. Selection of upmixers

The two commercial systems used were Dolby Pro-Logic II (DPLII) and Circle Surround II (CSII). As with the ASUS, both DPLII and CSII are *natural spatialization algorithms* [29] as they do not simply add reverberation to create the new channels but utilize spatial information already in the original signals. The ASUS system used was not the same as for that in the subjective experiment reported previously; it included a degree of cross-talk in the input signals (mixed at -5 dB), which reduced the output signal levels.

5.2. Method and stimuli

Besides the new upmixer [4], a commercially available surround-sound processer (manufactured by Marantz, model SR4400 "AV Surround Receiver") with a DPLII and CSII upmixer was used to create

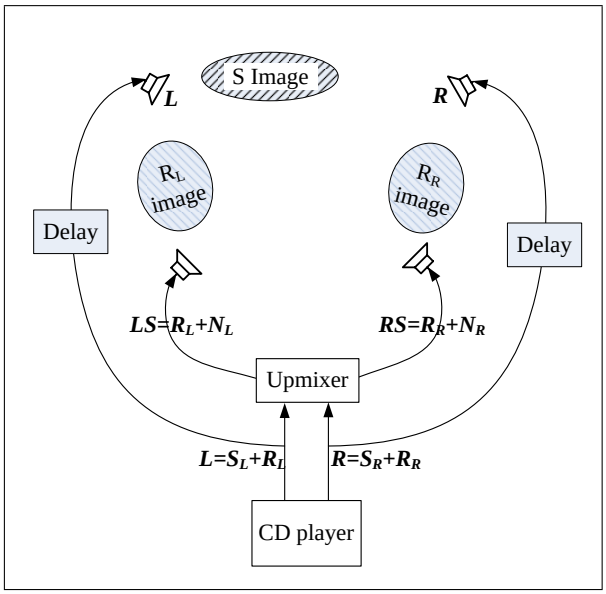


Fig. 1: Overview of source and reverberance imagery created by an upmix system according to the discussed design criteria. The system processes two input signals (L and R) of a musical recording (such as the output of a CD player) and creates two new signals LS and RS which are radiated by the rear loudspeakers. The front loudspeakers radiate the original signals with a time delay to account for the IO latency of the upmixer. 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 system tries to extract these reverberance image components and radiate them from the rear channels, creating reverberance phantom images to the side of the listener. Other "noise" artifacts created by the 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 channel; that is, signal RS should be uncorrelated with signal L, and LS uncorrelated with R.

the upmixed audio signals. The Marantz unit was fed a two-channel input signal and the five processed channel outputs were recorded. The processor was configured so that the subwoofer channel was not used. The "music mode" was selected for each upmixer, but there were no other options available (such as changing the time delay for the rear channels).

Three two-channel stimuli were used:

1. White noise reproduced from a single loudspeaker in a medium-sized concert hall. The recording was made using a forward-facing spaced microphone pair (50 cm spacing, DPA type 4011 cardioid microphone), 3.5 metres from the source, with the source equidistant to each microphone.

2. 40 second anechoic recording of sung female voice reproduced from loudspeaker and recorded in the same way as for the noise source, except the loudspeaker was 3 metres off-axis (in the direction of stage-right).

3. 23 second two-channel recording of *Her Majesty* by The Beatles, from a CD. This was chosen because it contains hard amplitude-panned sources, as an example of a mixing technique common in pop-music. A time-domain plot is shown in figure 2 and lissajous plots in figure 3 which shows that the source (a guitar) is hard-panned in the right channel (region B), and slowly moves to the left channel.

5.3. Electronic analysis of signals

Two electronic properties of the output signals for the three upmixers were investigated: signal level and inter-channel correlation. The first was measured as a level ratio between either the front and rear loudspeaker signals (i.e. the dB level of each front channel signal was logarithmically summed) or the ratio of the rear-right channel level relative to the front-right channel level. These level analyses are summarised in figure 4. Secondly, the correlation between the output channels was analysed by calculating the average cross-correlation in a 23 ms window between a variety of signal pairs[2]; as summarised in figure 5 for a noise signal and figure

[2]For two signals, say the L and R loudspeaker channel vectors, the cross-correlation was calculated using the following MATLAB incantation: `xcorr_LR=xcorr(L,R,1024,'coeff')`.

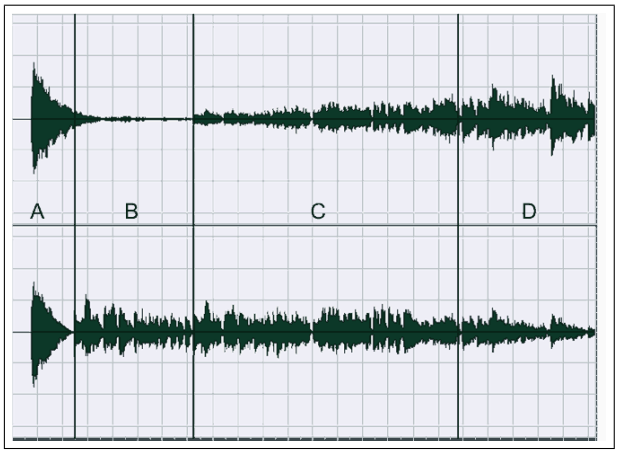
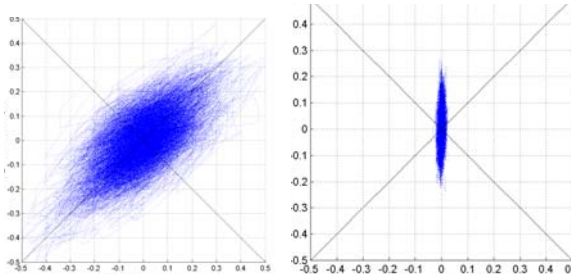
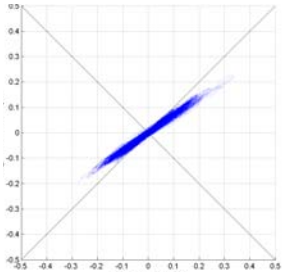


Fig. 2: Time-domain plot of a recording which contains both hard-panned music and dynamic amplitude panning: *Her Majesty* by The Beatles (23 seconds long).



(a) Region A (b) Region B



(c) Region D

Fig. 3: Lissajous phase plots for various regions of *Her Majesty*. x axis is the left channel, and y axis is the right channel (voltage).

6 for the voice signal- both recordings made in the concert hall as described.

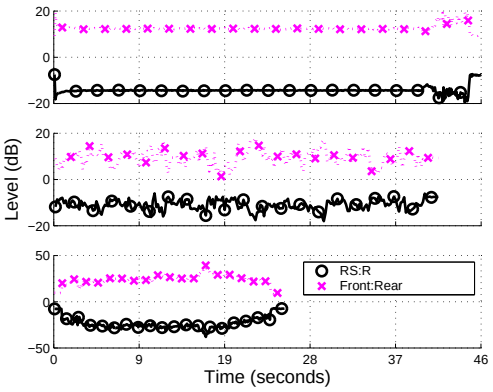
Looking at the level analysis in figure 4, we see that for the off-axis voice recordings (the middle sub-plots), the rear loudspeaker signal level is high for all upmixers. For the DPLII system, there are occasions when the front-loudspeaker level is actually *less* than the rear. When this occurred, the source image was localized in the rear loudspeakers as well as the front, which seemed unnatural (there was no front-stage). This is not the case with the ASUS, which compensates for time delay panning using an adaptive filter to automatically time-align the input signals before the difference (i.e. "ambiance") signal is calculated [4].

Regarding the *Her Majesty* piece, it can be seen that the DPLII system can not cancel the hard-panned source and it appears in the RS channel with a high level. The new ASUS upmixer accounts for "hard" amplitude panning by introducing cross-talk in the input signals so that a signal in only one channel can be canceled (otherwise it would appear in the front and rear channels).

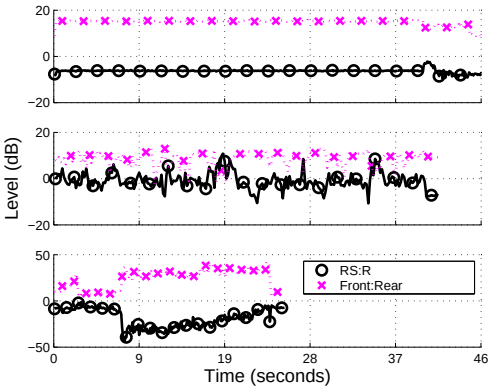
The time-averaged cross-correlation measurement undertaken is not ideal, as variation in the **IACC** over time have been shown to affect the width of source images [30]. However, it gives a basic insight into how the output channels of the upmixers maybe be perceived. As mentioned in the subjective design criteria: *Spatial distortion of source image (compared with 2/0 loudspeaker audition) should be minimized.* In the electronic design criteria in 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***, it is seen that these signals are quite correlated for the CSII and DPLII systems, yet these signals are uncorrelated for the ASUS system. Comparing the left and right signal correlation does not give a representative idea about how the source image might be perceived with the different systems, as DPLII and CSII both produce a centre loudspeaker channel as well.

Looking at the correlation between the side loudspeaker channels (i.e $R-RS$), we can see that the rear channels for the CSII are delayed by 440 sample- i.e. 10 ms. (Although some DPLII systems also have a delay for the rear channels, this is not recommended for the "music mode" [19].) For the rear loudspeaker channels, the correlation was unity for the CSII: in other words they were identical; a "mono" surround. Listening to this it was quite obvious; the reverberance seemed to decay to a point directly behind the listener, which was quite unnatural and at times even irritating (although listening to one channel alone, the temporal structure of the reverberance seemed quite natural and pleasant). For the DPLII and ASUS systems there was a negative correlation between the rear loudspeaker channels ($LS-RS$). This negative correlation was sometimes noticeable with loud transient attacks (a negative correlation for front loudspeaker signals produces a close-sounding source image; [26]). However, the reverberance image sounded very natural and its timbral quality was robust for off-axis listening positions; in contrast to upmixers which use comb/allpass filters for decorrelating the rear channel (e.g. [20]).

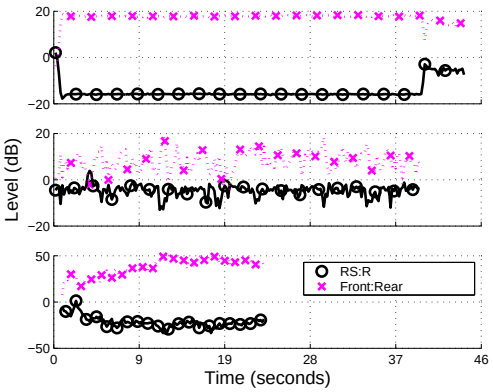
5.3.1. Output signal level



(a) CSII



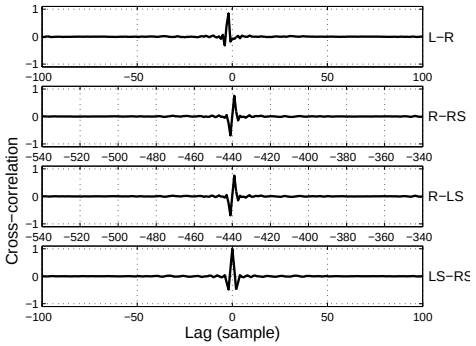
(b) DPLII



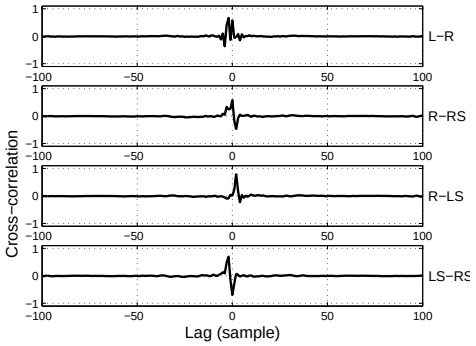
(c) ASUS

Fig. 4: Energy ratio of front:rear channels and RS to R for different upmix systems with three different two-channel input signals. In each subplot: Top plot is for noise source; middle plot is for off-axis voice recording; bottom plot is for *Her Majesty* by The Beatles.

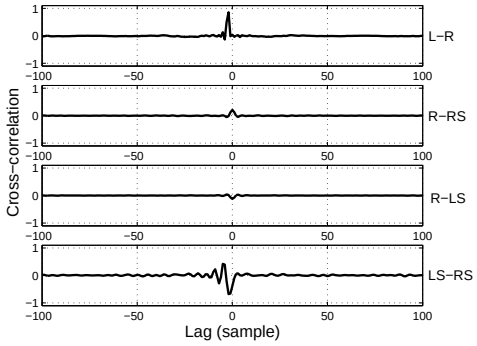
5.3.2. Output signal correlation



(a) CSII

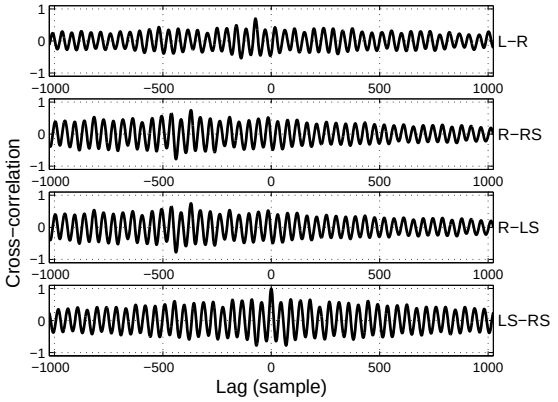


(b) DPLII

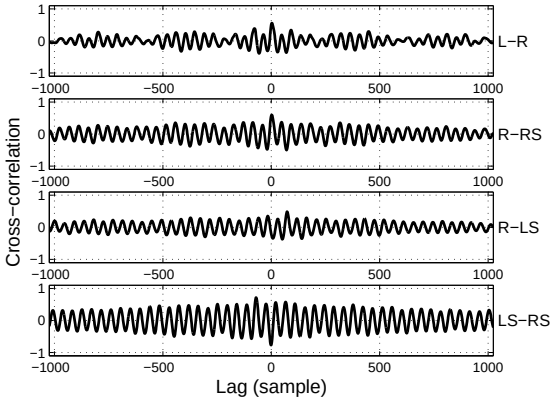


(c) ASUS

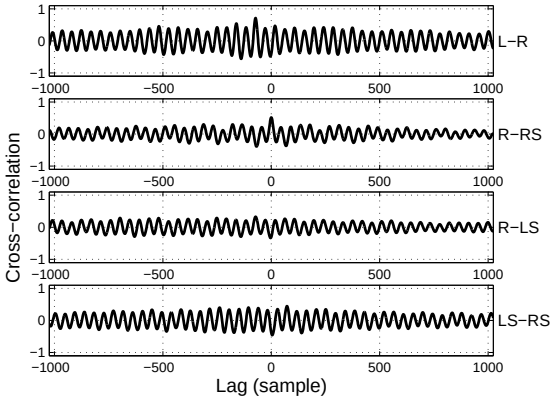
Fig. 5: Cross-correlation between front, side and rear channel output signals of different upmix systems (e.g. the top subplot L-R shows the cross-correlation between Left and Right output signals). Two-mike recording of white noise reproduced with a loudspeaker in concert hall. Note different lag range for CSII: this is due to a 10 ms delay of the rear speakers. Each graph is zoomed in around the main peak. The noise source was slightly off-axis, hence the main peak occurs at a lag of -4 samples.



(a) CSII



(b) DPLII



(c) ASUS

Fig. 6: As figure 5 but with the voice source instead of noise source.

6. CONCLUSION

The context of this paper is upmixing of unencoded "off the shelf" stereo music recordings, for reproduction with four or five surrounding loudspeakers. Subjective and objective evaluation methods are suggested for ensuring that the upmixed sound scene maintains a spatial image fidelity which is true to the mixing intentions of the recording engineer. This is qualified in terms of perceived spatial properties of imagery relating to both the recorded sound source (source imagery) and imagery related to the recording environment (ambiance or reverberance imagery), as well as electronic analysis of the correlations and levels of the output signals.

The underlying premise for the work is that a high-quality upmixer should leave source imagery spatially undistorted by the signal processing, and that reverberance imagery be distributed between the side loudspeakers so as to maintain a sense of lateral reverberance. A method using a graphical mapping device was suggested for investigating both spatial distortion of source imagery due to an upmixer and reverberance image distribution, and empirical data from a study with a new upmixer was analysed to show how performance with respect to these criteria can be evaluated. Three upmixers were also compared with respect to the electronic design criteria, which shows how time-delay and amplitude panning can reduce the efficacy of some commercial upmixers in terms of extracting sound components which affect reverberance imagery, and lead to undesirable artifacts.

7. ACKNOWLEDGEMENTS

This work was sponsored by Bang and Olufsen and a grant from the National Sciences and Engineering Research Council of Canada. The electronic measurements were undertaken at Philips Research, and the author thanks Dr. R. Aarts for his advice and assistance with this.

References

[1] F. Rumsey. Spatial audio and sensory evaluation techniques- context, history and aims. In *Proceedings of Spatial audio and sensory evaluation techniques conference*, Guilford, UK, 2006.

[2] 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.

[3] T. Letowsky. Sound quality assessment: concepts and criteria. In *Proceedings of the AES 87th international convention*, New York, 1989.

[4] J. S. Usher. *Subjective evaluation and electroacoustic validation of a new approach to audio upmixing using adaptive filters.* PhD thesis, McGill University, Schulich school of music, 2006 (submitted).

[5] 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.

[6] 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.

[7] A.H. Marshal and M. Barron. Spatial responsiveness in concert halls and the origins of spatial impression. *Applied Acoustics*, 62:91–108, 2001.

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

[9] 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.

[10] S. Bech. Methods for the identification of primary subjective attributes of spatial sound quality. In *Proceedings of the 2001 International Workshop on Spatial Media*, Aizu-Wakamatsu, Japan, October 2001.

[11] N. Zacharov and K. Koivuniemi. Audio descriptive analysis and mapping of spatial sound displays. In *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, 2001.

[12] 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.

[13] D. H. Mershon. Phenomenal geometry and the measurement of perceived auditory distance. In R. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 13, pages 257–274. Erlbaum, Mahwah, NJ, 1997.

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

[15] 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.

[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] K. Gundry. A new active matrix decoder for surround sound. In *Proceedings of the*

AES 19th international conference : Surround Sound - Techniques, Technology, and Perception, 2001.

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

[19] R. Dressler. Dolby Surround Pro Logic II Decoder Principles of Operation. Dolby Laboratories Information, 2000.

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

[21] J. S. Abel and D. P. Berners. Reverberation acoustics, analysis and synthesis. *Tutorial session T21, 117th Convention of the AES, San Francisco*, 2004.

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

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

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

[25] 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.

[26] W. L. Martens. The impact of decorrelated low-frequency reproduction on auditory spatial imagery: Are two subwoofers better than one? In *Proceedings of the AES 16th international conference on spatial sound reproduction*, pages 87–77, Rovaniemi, Finland, 1999.

[27] F. Jacobsen and T. Roisin. The coherence of reverberant sound fields. *Journal of the Acoustical Society of America*, 108(1):204–210, 2000.

[28] S. Kim, W. L. Martens, and A. Marui. Discrimination of auditory source focus for musical instrument sounds with varying low-frequency cross correlation in multichannel loudspeaker reproduction. In *Proceedings of the AES 119th international convention*, 2005.

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

[30] 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.