2009-02-01

Localization Quality Assessment in Source Separation-Based Upmixing Algorithms

Dan Barry
Technological University Dublin, dan.barry@tudublin.ie

Gavin Kearney
Trinity College Dublin, gpkearney@ee.tcd.ie

Funder: Enterprise Ireland; Science Foundation Ireland

LOCALISATION QUALITY ASSESSMENT IN SOURCE SEPARATION BASED UPMIXING ALGORITHMS

DAN BARRY[1] AND GAVIN KEARNEY[2]

[1] *Audio Research Group, Dublin Institute of Technology, Ireland*
dan.barry@dit.ie
[2] *Music and Media Technologies, Trinity College, Dublin*
gpkearney@ee.tcd.ie

In this paper we explore the source localisation accuracy and perceived spatial distortion of a source separation based upmix algorithm for 2 to 5 channel conversion. Unlike traditional upmixing techniques, source separation based techniques allow individual sources to be separated from the mixture and repositioned independently within the surround sound field. Generally, spectral artefacts and source interference generated during the source separation process are masked when the upmixed sound field is presented in its entirety; however, this can lead to perceived spatial distortion and ambiguous source localisation. Here, we use subjective testing to compare the localisation perceived on a purposely generated discrete presentation and an upmix (2 to 5 channel) of the same source material using a source separation based upmix algorithm.

INTRODUCTION

Surround Sound technology has become common place in modern gaming and entertainment applications. Whilst a large proportion of audio content is authored specifically for multi-channel reproduction, some pre-existing content is often repurposed for surround sound presentation. Upmixing techniques are typically used to generate several reproduction channels from a limited number of source channels. Traditional approaches often involve ambiance extraction, typically through mid-side processing, and channel delay schemes to increase immersion in the resultant sound field. Although these approaches do provide a greater sense of spatialisation, they do not facilitate localisation of discrete sound sources within the surround sound field. Upmixing techniques based on sound source separation algorithms afford the possibility of repositioning sources discretely within the surround field offering greater upmix flexibility.

This study is not concerned with comparing existing separation algorithms for the purposes of upmixing, rather, the purpose of the experiment proposed here, is to subjectively compare the localisation perceived on a purposely generated 5 channel presentation and an upmix of the same source material using a source separation based upmix algorithm. Purpose generated multi-track recordings are used to create both a 5 channel mix and a 2 channel mix. Using the source separation based upmix algorithm, the 2 channel mix is then upmixed to emulate the discrete 5 channel mix. Using subjective testing, it is then possible to directly compare the localisation achievable between the purpose generated 5 channel mix and that of the 2 channel upmix. For the experiments we use a modification of the ADRess algorithm [6] as the basis for our upmixing model. The algorithm uses a novel spatial clustering and adaptive filtering technique to identify and separate sources in real-time based on their location within the stereo field. The sources can then be remixed and/or re-authored with relative ease.

1 BACKGROUND

1.1 Traditional Upmixing Techniques

The origin of up/down-mixing techniques can be traced back as far as the Quadraphonic era, where four discrete channels of audio were encoded onto two channel vinyl discs [1]. The discs accommodated playback on standard stereophonic record players or four channel playback with dedicated Quadraphonic decoders. Unfortunately, due to competing technologies, increased production costs, and a confused public, the Quadraphonic era ended in a complete commercial failure.

However, by the end of its demise, the principles of 'matrix' encoding and decoding on which Quadraphonics was founded had already migrated from the domestic environment to the cinematic world. In 1975, Dolby Systems introduced 'Dolby Stereo' [2], a method of encoding four cinematic audio channels onto the two optical channels found at the side of 35mm cinematic film. The original studio master reproduction channels, *L*, *R*, *C*, and *S* (the left, right, centre and surround channels respectively) are encoded onto the L_T and R_T channels of the optical soundtrack. Decoding of the *S* and *C* channels involves the sum and difference of the two optical L_T and R_T channels, such that phase shifted surround components will cancel each other out in the decoded centre channel, and that the centre

channel will be removed from the decoded surround channel. This is achieved by several matrix operations as outlined in [3].

A major consequence of such matrixing is the crosstalk inherent in each channel. Both the surround and centre channel components in the decoded L_{Front} channel are each only 3dB down from the original L component. This is the same for the R_{Front} channel. Crosstalk in the surround channel is overcome by delaying the surround feed such that localisation precedence is maintained towards the three frontal channels. Pro-Logic, the consumer version of Dolby Stereo, improves image stability somewhat by including active 'logic steering' circuitry which attempts to steer images towards one speaker. The control circuit looks at the relative levels and phases of the input signals in order to control a group of VCAs which govern the antiphase signals in the output matrix. However, in a 5 speaker setup, the VCAs do not control steering in the Left-Right axis and the Front-Back axis separately. In Pro-Logic II [4], each axis operates individually through inclusion of a feedback servo control system that adjusts the levels of the VCAs controlling the L_T, R_T, L_T+R_T and L_T-R_T signals such that better channel separation can be achieved.

Such matrix encoding and decoding has received marketplace acceptance as the standard for cinematic upmixing, but we must bear in mind that the majority of stereophonic *music* presentations are not matrix encoded. This leads to distinct differences between how Pro-Logic systems handle cinematic and music program material. Music mode in Pro-Logic II systems includes a high-shelf filter in the surround channels, whereas movie mode does not. There is also no delay component for the rear channels, which although desirable for coincident arrival wavefronts at the centre listening position (in particular transients), can lead to a perceived reduction in channel separation.

It is clear that although matrix systems have significantly developed from their beginnings as humble passive decoders into sophisticated solutions for up-mixing from two-channel material, their application to all types of program material is not fully satisfactory. Furthermore, the fact remains, that in order to obtain optimal performance from any matrix system, the two channel material needs be properly preconditioned (encoded) beforehand.

1.2 Source Separation and Upmixing

Sound source separation refers to the task of extracting individual sound sources from some number of mixtures of those sound sources. Unlike matrixing technology, the source material does not have to be pre-encoded for effective upmixing to be achieved. In recent years, advances in dual channel sound source separation technology such as the DUET algorithm [5] and the ADRess algorithm [6] have made it possible to achieve high quality separation of individual sources from stereophonic mixtures. The former is applicable for speech separation in spaced sensor convolutive mixtures whereas the latter is designed for separating or 'de-mixing' intensity panned (linear mixed) stereophonic music content. The primary focus in development and application of [5] and [6] above was purely that of sound source separation. However, prior to [6], the application of similar techniques specifically for the purposes of upmixing had been developed in Creative Labs [7] where it was shown that the use of weighted time-frequency masking could be applied effectively in multi-channel upmixing. More recently, the same algorithms have been applied to upmixing for Wave Field Synthesis applications [8].

It has been shown in the past that these algorithms are capable of adequate source separation but at the cost of both temporal and spectral artefacts when the sources are reproduced in isolation. Objective comparisons of a number of source separation algorithms are presented in [9] and [12]. In general however, such artefacts are perceptually masked when the sound field is reconstructed even after manipulation of individual sources. However, if the content is repurposed for surround presentations, the same artefacts can theoretically manifest themselves through spatial distortion and localisation ambiguity. This can be appreciated if one considers that using the aforementioned separation algorithms; a separated source will often contain time varying interference from overlapping sources within the mix. When the separated sources are then relocated in a multi-channel presentation, this interference becomes apparent as channel crosstalk which inherently leads to image shifts in the surround field. The purpose of this paper is to explore the subjective effects of this image shifting by directly comparing a discrete 5 channel mix and an upmix of the same material.

2 UPMIXNG MODEL

For this experiment we use the ADRess algorithm [6] with the addition of an azimuth windowing function which was suggested in [7]. The ADRess algorithm achieves source separation by taking advantage of destructive phase cancellation in the frequency domain. For each frame, m, of a short-time Fourier representation of the signal, one channel is iteratively gain scaled and subtracted from the other in the complex frequency domain after which the absolute value is taken. The resulting array is of dimension N x $ß$, where N is the number of frequency points and $ß$, the

azimuth resolution, is the number of equally spaced gain scalars between 0 and 1. The operation reveals local minima, due to phase cancellation across the azimuth plane for each frequency component. Using a simple clustering technique, components belonging to a single source are seen to have their minima in a localised region about some gain scalar which ultimately refers to the intensity ratio between each channel, i.e., the pan position of the source in stereo space. By estimating the magnitude of each of the time-frequency minima and only resynthesising those with a desired intensity ratio, a single source maybe reconstructed. The original mixture phase information maybe used as was shown in [10]. The process can be summarised as follows with the iterative gain scaling process achieved using equation (1) where $X_j(k,m)$ is a complex frequency domain representation of the m^{th} frame of the j^{th} channel (left or right).

$$\begin{aligned} Az_1(k,m,i) &= \left| X_2(k,m) - g(i)X_1(k,m) \right| \\ Az_2(k,m,i) &= \left| X_1(k,m) - g(i)X_2(k,m) \right| \end{aligned} \quad (1)$$

where $1 \leq k \leq N$, N being the Fourier transform length, and where $g(i) = i/\beta$, for all i where, $0 \leq i \leq ß$, and where i and $ß$ are integer values. $ß$ refers to the number of gain scalars to be used and ultimately gives rise to the resolution achieved in the azimuth plane. The resulting matrix $Az_j(k,m,i)$ represents the frequency-azimuth plane for the m^{th} frame of the j^{th} channel. Each of k frequency bins will exhibit a local minimum at some index i. It can be observed that the majority of frequency bins pertaining to a single source should exhibit their minima around a singular value for i. These local minima represent the points at which frequency components experience a reduction in energy due to destructive phase cancellation between the left and right channel. This energy reduction is directly proportional to the amount of energy which the cancelled source had contributed to the overall mixture and so to invert these minima around a single azimuth point should yield short-time magnitude spectra of the individual sources. To achieve this inversion, we simply subtract the minimum from the maximum of the function in (1) for each of k frequency bins as described in equation (2).

$$\bar{Az}_1(k,m,i) = \begin{cases} Az_1(k,m)_{\max} - Az_1(k,m)_{\min} & \text{if } \bar{Az}_1(k,m,i) = \min \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where 'min' and 'max' refer to the global minimum and maximum of the k^{th} frequency-azimuth function. Note that the inverted frequency-azimuth plane for channel 2 is created in an identical fashion. Now, the instantaneous magnitude spectrum of a single source or subspace at pan position d, predominant in the j^{th} channel can be approximated as in (3)

$$Y(k,m) = \sum_{i=d-H/2}^{i=d+H/2} \bar{Az}_j(k,m,i) \times \left(1 - \frac{2|d-i|}{H} \right) \quad (3)$$

where d is the azimuth index, i.e. the pan position of the source for separation and H is the azimuth subspace width which is simply a neighbourhood around the azimuth index. The second term in (3) simply creates a linear weighting function such that components further from the azimuth index are scaled down. This essentially creates a triangular separation window along the azimuth axis. As we will see, the properties of this window will allow adjacent azimuth subspaces to be overlapped in such a way as to allow the extraction of, in this case, 5 discrete subspaces for surround presentation. *YR(k)* is now an N x 1 array containing the short-time magnitude spectrum of a single source or azimuth subspace. For a detailed description of the ADRess algorithm, refer to [6].

3 OBJECTIVE TESTING

Although the algorithms described here and in [5] and [7] are capable of perceptually acceptable separations, a certain degree of signal interference from other sources in the mixture is inevitable in each separation. This section describes the theoretical errors which are known to occur in such algorithms. The material objectively evaluated here is the same as that used for subjective testing in section 4.

In the case of the algorithm described above and used in this experiment, increasing the value of H will result in capturing more of the desired source for resynthesis but will also lead to a lower signal to interference ratio due to time-frequency (TF) overlap between sources. Theoretically, if the sources do not exhibit TF overlap, near perfect recovery of all sources is possible. However, where western tonal music is concerned, a significant amount of overlap can be assumed. Given that equations (1) and (2) use both phase and magnitude information to estimate the location of each TF point, the inherent TF overlap between sources causes the local minima to spread out from the true source locations. This is referred to as frequency azimuth smearing in [6]. This can be observed in Figure 1, where the inverted frequency-azimuth plane (N=4096, β=100) for a single frame of the stereo audio is shown. The audio used here is described in greater detail in section 4.1. The audio frame contains 5 sources (guitar, bass, drums, vocals and piano) distributed equally across the stereo field. Referring to Figure 1, each frequency component has been resolved to a location within the stereo field. Components naturally cluster close to the

theoretical source locations but it can be seen that some components are incorrectly localised and so wider subspace widths (*H*) would be required to faithfully approximate sources at the cost of unwanted interference.

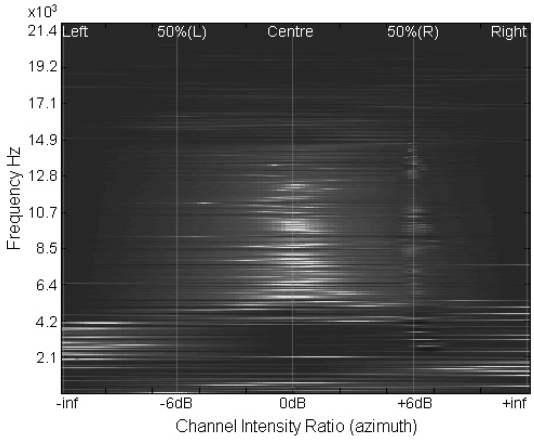

Figure 1: Inverted frequency-azimuth plane for a single audio frame as described by equation 3. Five sources are clearly present, distributed equally from far left (-inf) to far right (+inf). Note the smearing of frequency components across the azimuth plane.

This ultimately means that the source estimates, $\hat{S}_j(t)$, are not equal to the true sources $S_j(t)$ but the sum of the source estimates should be approximately equal to the sum of the true sources as in (4).

$$\hat{S}_j(t) \neq S_j(t) \quad \text{but...} \quad \sum_{j=1}^{J} \hat{S}_j(t) \approx \sum_{j=1}^{J} S_j(t) \tag{4}$$

This is a known shortcoming of such separation algorithms. Nevertheless, in the case where the stereo presentation is reconstructed, even with individual source manipulation, the artifacts are generally not discernable [11] but the same artefacts could theoretically lead to noticeable localisation ambiguity when reproduced for surround presentation. Section 4 explores this issue further.

3.1 Reconstruction Errors

The frequency-azimuth smearing illustrated in Figure 1 essentially leads to reconstruction errors in each of the individual source estimates. This reconstruction error will depend ultimately on the number of instantaneously active sources and their relative TF overlap. In [12], a set of objective measurement criteria were presented in order to compare the reconstruction quality of a number of source separation algorithms. The criteria proposed were as follows:

ISR – Image to Spatial distortion Ratio (dB)
This measurement assesses the algorithms ability to estimate the individual source contributions to each channel in the mixture signal.

SIR – Source to Interference Ratio (dB)
Here, the presence of unwanted interference from other sources in the mixture is measured as a function of the source estimate itself.

SAR – Source to Artifact Ratio (dB)
Additional algorithm specific artifacts are also measured as a function of the source estimates.

SDR – Signal to Distortion Ratio (dB)
This measurement conveniently combines all error measurements described above. Refer to [12] for a detailed description of the derivation of these measures.

In order to have some objective measures to refer to for comparison purposes, the subjective test material used in section 4 has been processed using the blind source separation evaluation toolbox [13] which implements the error measurements described above. Figure 2 presents the error measurement criteria for each of 5 source estimates separated from the stereo mix. These 5 source estimates will ultimately comprise the 5 channel upmix in section 4. Note, the original implementation uses the $10\log_{10}$ power law for error measurement but here we use the $20\log_{10}$ power law given its prevalence in the audio domain.

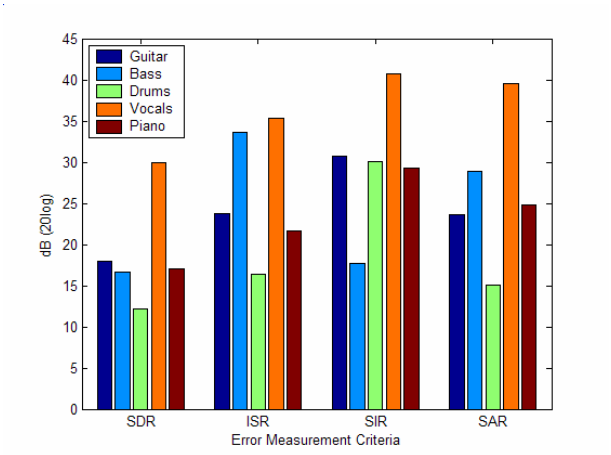

Figure 2: SDR, ISR, SIR and SAR for each of the five separated sources from stereo mixture from which the experimental upmix will be generated. Sources positioned from far left to far right as follows: guitar, bass, drums, vocals and piano.

Referring to Figure 2, it can be seen that the vocal has achieved the greatest amount of separation owing to the fact that it is the most prevalent source in the stereo mix. Subsequently, the bass, the lowest source in the stereo mix achieves the poorest SIR. This is a property of almost all separation algorithms, whereby the loudest sources will generally have greatest influence during clustering stages. Both guitar and piano exhibit similar

error values owing to the fact that they exhibit significant TF overlap (between each other) and are of similar amplitude in the stereo mix. In general however, it can be seen that in this example, an average SIR of 30dB can be achieved with a minimum of 17dB in the case of the bass.

3.2 Image Shifting

Given that source separation is generally the task of solving an under-determined problem, theoretical errors are inevitable as discussed above. As such, we consider the effects of such errors when separation algorithms are used for multi-channel upmix. As described above, interference from nearby sources is the most prevalent problem, whereby an individual source estimate will invariably contain some unwanted components from other sources. Consider the upmix task, where in this case 5 virtual sources from the stereo mixture will be repurposed as 5 discrete sources for a 5 channel presentation. This source interference becomes channel crosstalk which should theoretically result in image shifting within the surround presentation. Subjectively, this should lead to localisation errors.

In order to illustrate how TF overlap causes localisation errors in the separation algorithm we derive the azimugram (time-azimuth representation) of the stereo mix used for upmixing in this experiment. Essentially each column in Figure 3 is the transposed column sum of a frame such as that presented in Figure 2. Referring to Figure 3, note the encircled area, where it can be clearly seen that source overlap has caused the source image to temporarily shift towards the centre. This theoretical error will result in channel crosstalk in any subsequent upmix of the material.

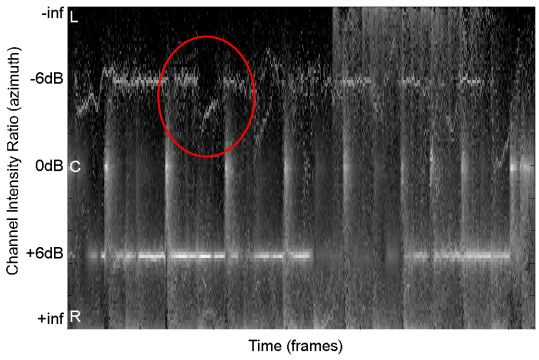

Figure 3: The time-azimuth representation of several hundred audio frames. Source activity is clearly visible as is source overlap leading to localisation errors in the source separation algorithm.

In the context of this experiment, we would expect SIR and ISR to be the most useful indicators of spatial distortion in the 5 channel upmix of the source material.

4 SUBJECTIVE TESTING

A subjective experiment was designed to compare the localisation accuracy of a 5 channel musical presentation created from an upmix using ADRess against a discrete 5 channel presentation. The aim of this test was to quantify the extent of localisation shifts due to the source interference in the upmixing algorithm. The test was performed in accordance with the ITU BS.1284-1 recommendations for listening tests [14] and conducted on a standard ITU 5-channel layout. Bass management (where low-frequency content from the main surround channels is routed to a subwoofer) was omitted from this experiment on the grounds that it may bias localisation of lower range sources.

4.1 Material Preparation and Stereo Mix

For the tests, a dedicated 2 channel stereophonic recording of a jazz ensemble was created. The recording consisted of 5 discretely recorded sources; Piano, drums, vocals, electric guitar and bass. The recordings are of studio quality and were taken at 96kHz, 16-bit. A stereo mix of the sources was generated such that the 5 sources were distributed equally across the stereo stage giving 5 equal width source subspaces that could be separated to produce the 5 channel upmix. The mixing criteria for the stereo mix is shown in Table 1.

Instrument	Level (rms)	Pan Position
Guitar	-5.8 dB	Left (100%)
Bass	-8.7 dB	Left (50%)
Drums	-7.2 dB	Centre
Vocals	0 dB	Right (50%)
Piano	-6.4 dB	Right (100%)

Table 1: Mixing parameters for stereo mix. Level measurements are normalised and averaged over 200mS frames where all 5 sources are present simultaneously.

The spectral contribution and relative mix intensity of each source can be seen in Figure 4. The drums are the most spectrally dense source, whilst the vocals contain the most significant energy in the mix. The bass guitar has the most limited frequency range with prominent spectral components below 300Hz.

4.2 Upmixing

In any 5 channel upmix, there are two-main methods of placing the audio sources: These are ‘audience-view’ (where the sources are kept at the front of the surround array and the rear speakers are used for lateral spatial enhancement), and ‘ensemble view’ (where the listener is put in the centre of the musical presentation, surrounded by the musical sources). The first approach is akin to ambience extraction, which is not the focus of this work. Here we adopt the latter approach, where we attempt to separate 5 equal width, overlapping, azimuth

subspaces from the stereo field (see Figure 5) so that each source might be uniquely mapped to a single loudspeaker in the 5 channel upmix. The modified ADRess algorithm described in section 2 was used for this purpose.

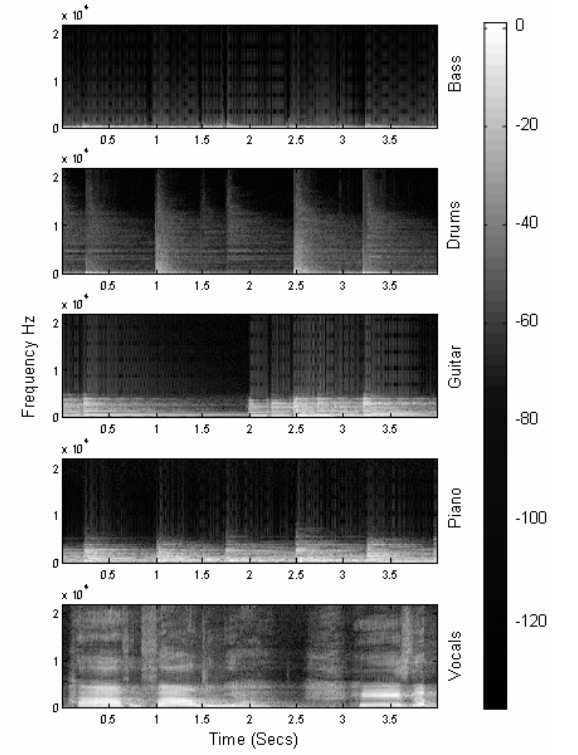

Figure 4: Spectrograms of discrete source contributions over 5 seconds of the two channel mix.

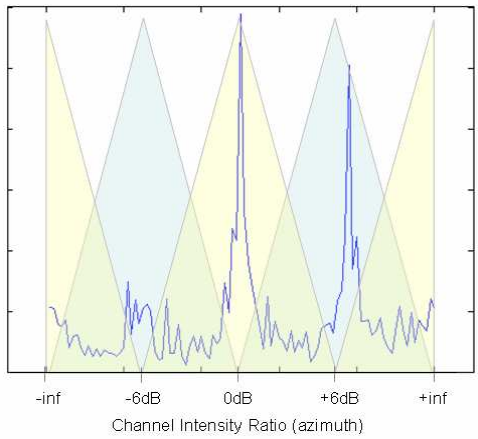

Figure 5: Stereo energy histogram illustrating the energy distribution across the stereo field from left (-inf) to right(+inf) within the stereo mix. ADRess is configured to separate 5 equal width overlapped subspaces for upmixing purposes.

4.3 Experimental Procedure

It was the task of each participant to attempt to identify the direction of the upmixed sources. For the upmix, there are 120 possible permutations by which all 5 sources can be mapped to the loudspeakers. However, we can limit the number of tests such that we are only interested in permutations where we can test localisation of each source uniquely mapped to each loudspeaker. Thus we only need to construct 25 different tests. This can be further reduced if we consider the symmetry of the array, since symmetrically equivalent tests should give identical results. This results in 15 unique tests with which to describe the localisation accuracy of the upmix. Also, for each upmix, there is then an exact discrete channel mix with which to compare the localisation accuracy, giving a total of 30 localisation tests for each participant.

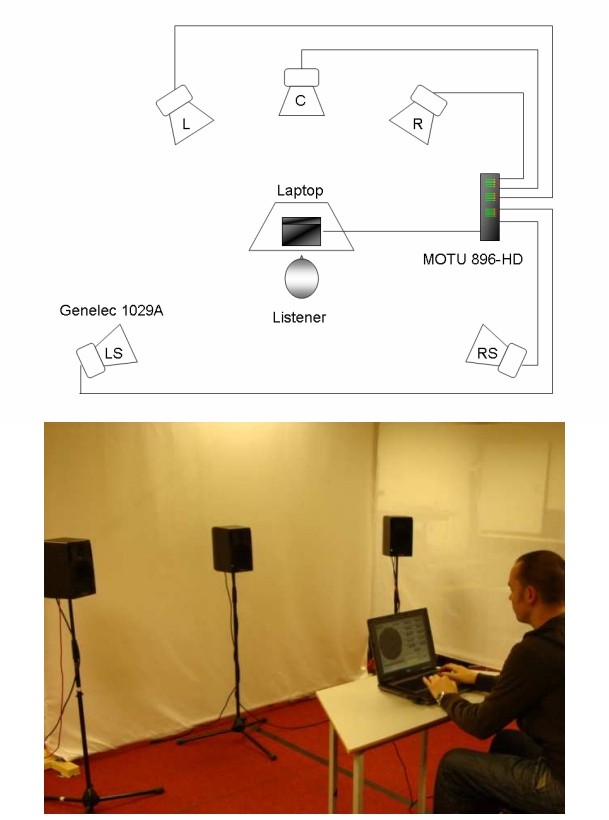

Figure 6: Listening Test Configuration. Bottom: Participant in the listening environment conducting the perceptual experiment with dedicated test software.

In total, 10 listeners were chosen for the tests, each under 35 years of age, of excellent hearing, and well experienced in musical production. The setup illustrated in Figure 6 consists of 5 Genelec 1029A loudspeakers each calibrated to 79dBC at the centre listening position. A MOTU 896-HD audio interface was used to route the audio to each of the loudspeakers and the test was controlled by the participant via a PC laptop. The

listening room is a good monitoring environment with a spatially averaged reverberation time of 0.3 seconds at 1kHz.

4.4 Data Acquisition

A dedicated software pointer, shown in Figure 7 was developed to perform the tests. The software gave each participant complete control over the test, allowing them to initiate the audio, stop the presentation or move on to the next presentation. For each test, the software asks the subject to identify the direction of one of the musical sources (shown in large yellow letters). The user can play the test presentation as many times as they desire, before they decide on the direction of localisation using the software pointer. The pointing tool consists of a circle displaying the ITU 5 channel layout with a moveable blue ball for choosing the source orientation. Given the diameter of the ball, there is a 1^{o} margin of error in the test software and the loudspeaker markers are $+/-3^{o}$ wide. The sequence in which each of the 30 samples is played is completely random and different for each participant.

Figure 7: Custom software designed for listening test.

The test results were compiled and are presented in the following section.

5 RESULTS

Observing the results of the subjective testing, it is apparent that the theoretical reconstruction errors discussed in section 3.2 have manifested themselves as image shifts within the upmix reproduction. This leads to localisation errors during subjective audition. However, the magnitudes of the errors are dependent on both the instrument and the channel in which it is reproduced. Firstly, we present the data for each reproduction channel (or symmetric pair) as the localisation error from the theoretical source position for each instrument in both the upmix and the discrete mix. Figure 8, 9 and 10 illustrate the perceived localisation error for the center, left/right, and left/right surround channels respectively. Both the discrete 5 channel mix and upmix errors are presented for comparison purposes. Note that 0 degrees refers to the normalised on axis angle for each reproduction channel.

5.1 Center Channel Localisation

Referring to Figure 8, it is apparent that the center channel localisation achievable within the upmix is largely similar to that of the discrete mix. Here, the mean localisation error is less than 5 degrees for drums guitar piano and vocals. The exception in both discrete and upmix presentations is the bass instrument, where a mean localisation error of 41 degrees and 25 degrees is apparent for the discrete mix and upmix respectively. In general, poor localisation of low frequency content is expected [15]. Note also that there is an image shift away from the discrete presentation toward the theoretical location. As a consideration, the SIR for the bass is poorest as indicated in Figure 2. This suggests that a substantial number of spectral components from the bass have 'leaked' into other separations. This of course translates to channel crosstalk in the upmix. Thus we postulate that in this case, the crosstalk has affected the perceived localisation of bass within the upmix to positive effect. The complex channel interactions could just as easily result in the opposite effect, shifting the source away from the intended location.

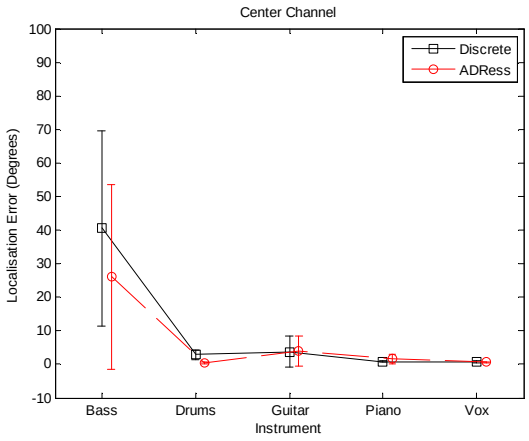

Figure 8: Perceived localisation deviations for discrete and upmixed sources positioned in the center channel with theoretical position 0 degrees. (95% Confidence Interval')

5.2 Left and Right Channel Localisation

Referring to Figure 9, for left and right channels a noticeable image shift is apparent between the discrete mix and the upmix. In this case, localisation achievable is clearly poorer for the upmix but the error remains below 10 degrees for drums, guitar, piano and vocals. The bass, as expected, achieves poorest localisation in both cases but a similar situation has occurred whereby the upmix image has been shifted toward the theoretical

source location. This has been discussed in the previous section. Note that the vocal has achieved the best localisation. This can be attributed to the fact that it was the loudest source in the stereo mix and achieved the greatest SIR (Figure 2) which inherently means that it will generate the least amount of crosstalk in the upmix leading to greater image stability.

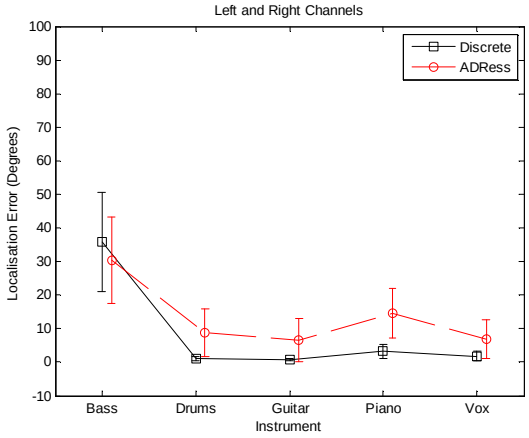

Figure 9: Perceived localisation deviations for discrete and upmixed sources positioned in the left or right channels with theoretical positions 30 degrees. (95% Confidence Interval')

5.3 Left and Right Surround Channel Localisation

In general, auditory events presented laterally to a listener are subject to the greatest localisation blur. Blauert [16] shows that sources presented to the sides of a listener undergo, on average, a localisation blur of +/- 10 degrees. Both the discrete and upmix presentations illustrate this trait. However, the upmix performs considerably poorer than the discrete mix for rear channels although the trend for each is similar.

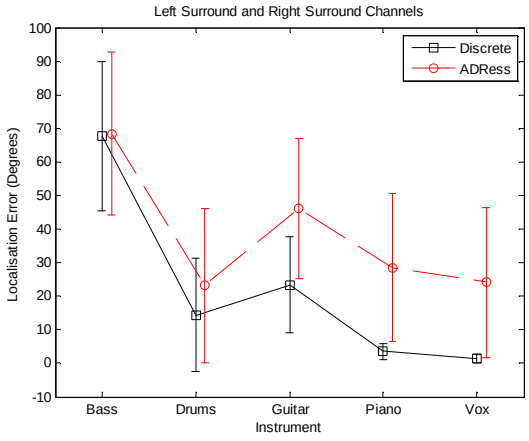

Figure 10: Perceived localisation deviations for discrete and upmixed sources positioned in the rear channels with theoretical positions 110 degrees. (95% Confidence Interval')

Note that on average, the upmixed images have shifted 40 degrees from the theoretical positions; however, the shift from the subjective discrete source locations is significantly less, in the region of 25 degrees on average. Given that the experiment is conducted in a real listening room as opposed to an anechoic chamber, the room acoustics impose constraints on the experiment. We therefore consider the discrete localisation results to be the ground truths as opposed to the theoretical source positions. With this in mind, Figure 11 presents the mean image shift of the upmixed source locations as a function of the discrete source locations.

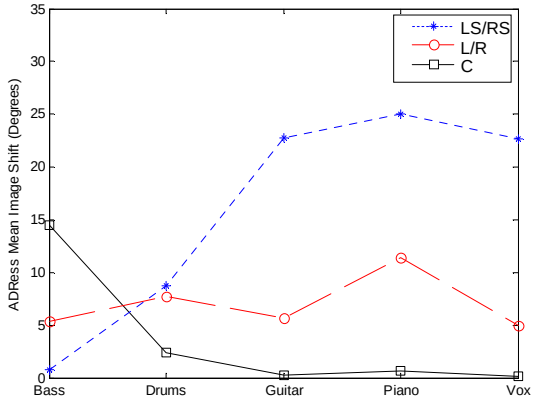

Figure 11: The mean image shift observed within the upmix material. (95% Confidence Interval')

5.4 Discussion

In general, the vocal has been localised most accurately in the upmixes with minimum image shifts in the frontal channels. Although the image shift from ground truth is considerable in the surround channels, it remains closer to the theoretical source position than other sources (Figure 10). Subsequently, the vocal also achieves the highest SIR (Figure 2) of all sources which implies that it will exhibit less crosstalk upon upmixing. This can be attributed to the fact that the source is almost 6dB louder than any other source in the mix which is advantageous for source separation. Referring to Figure 2, the drums achieve the poorest SIR but localisation accuracy remains strong in subjective testing. In general, transients are easier to localise due to the broadband nature of the instruments attack. Secondly, although the drums don't exhibit sustained loudness, they may frequently but briefly become the dominant source in the mixture upon their onset. This aids localisation and would inherently lead to a higher *instantaneous* SIR value. As discussed, bass is difficult to localise in most circumstances. This is evident in both the discrete and upmix presentations. In the case of piano and guitar, they achieve similar localisation accuracy with guitar localisation slightly outperforming

that of the piano. This is also supported by the objective measurements where the SIR for guitar is slightly better than that of piano.

In addition to localisation errors, some subjects noted, in rare cases, additional artifacts which were later attributed to upmixed material. Occasionally, some transients were perceived as 'dulled' with respect to the discrete mix although not objectionable. In general, however, many subjects reported that they were often unable to identify which of the two presentations they were listening to in a given test. Finally, it should be noted that in a real world scenario, the listener has no prior expectation of source locations and so localisation errors are not detrimental to the effective application of source separation to upmixing, provided that the artifacts known to exist in individual reproduction channels (separations) are masked when the full presentation is recreated.

6 CONCLUSIONS

In this paper, the source localisation accuracy and perceived spatial distortion of a source separation based upmix algorithm for 2 to 5 channel conversion was investigated. Subjective and objective testing methodologies were presented in order to assess the localisation accuracy. It was shown that theoretical reconstruction errors associated with the source separation process manifest themselves as image shifts in the upmix presentation and thus lead to perceived localisation distortion. However, the localisation error is acceptable in center, left and right channels but significant in the surround channels, yet still below 30 degrees. The tests carried out here are not intended to be comprehensive, but rather, indicative that separation algorithms are suitable for upmix applications, particularly for audience view/ensemble view conversion.

7 ACKNOWLEDGEMENTS

The contribution of the participants in the subjective listening tests is gratefully acknowledged by the authors. Dan Barry acknowledges the support of Enterprise Ireland. Gavin Kearney acknowledges the support of Science Foundation Ireland.

REFERENCES

[1] J. M. Eargle. "Multichannel Stereo Matrix Systems: An Overview". Journal of the Audio Engineering Society, 19(7):552-559, 1971.

[2] J. Hull, "Surround Sound Past, Present and Future", Dolby Laboratories Licensing Corporation, 1994.

[3] R. Dressier, "Dolby Pro Logic Surround Decoder Principles of Operation", Dolby Laboratories Licensing Corporation, 1993.

[4] "Dolby Surround Pro Logic II Decoder Principles of Operation", Dolby Laboratories Licensing Corporation, 2004.

[5] A. Jourjine, S. Rickard, O. Yilmaz, "Blind Separation of Disjoint Orthoganal Signals: Demixing N Sources from 2 Mixtures", *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, June 2000

[6] D. Barry, R. Lawlor and E. Coyle, "Sound Source Separation: Azimuth Discrimination and Resynthesis", *Proc. 7th International Conference on Digital Audio Effects, DAFX 04*, Naples, Italy, 2004

[7] C. Avendano, J.M. Jot, "Frequency-Domain Techniques for Stereo to Multichannel Upmix", *In Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland 2002. pp. 121-130,

[8] M. Cobos, J. J. Lopez, A. Gonzalez and J. Escolano, "Stereo to Wave-Field Synthesis Music Upmixing: An Objective and Subjective Evaluation", *In Proc. ISCCSP 2008*, Malta, 12-14 March 2008, pp. 1279 -1284

[9] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation", IEEE Transactions on Speech and Audio Processing, vol. 14 (4), pp.1462–1469, 2006

[10] D. Barry, R. Lawlor, and E. Coyle, "Comparison of Signal Reconstruction Methods for the Azimuth Discrimination and Resynthesis Algorithm", *Proc. 118th Audio Engineering Society Convention*, May 28-31, Barcelona, Spain, 2005

[11] C. Avendano, "Frequency Domain Source Identification and Manipulation In Stereo Mixes for Enhancement, Suppression and Re-Panning Applications", *In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 55-58 New Paltz, NY, October 19-22 2003

[12] E. Vincent, H. Sawada, P. Bofill, S. Makino and J.P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and

results", *in Proc. Int. Conf. on Independent Component Analysis and Signal Separation*, 2007.

[13] C. Févotte, R. Gribonval and E. Vincent, "A toolbox for performance measurement in (blind) source separation", http://bass-db.gforge.inria.fr, accessed November, 2008.

[14] The ITU Radiocommunication Assembly, Recommendation ITU-R BS.1284-1 "General methods for the subjective assessment of sound quality", 2002.

[15] G. Theile "On the localisation in the Superimposed Soundfield", Dissertation, Technische Universität Berlin, 1980.

[16] J. Blauert, 'Spatial Hearing', Revised Edition, MIT Press, 1996.