Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK



# Modification of Spatial Information in Coincident Pair Recordings

Jeremy Wells

Audio Lab, Department of Electronics, University of York, York, North Yorkshire, YO10 5DD, England
jjw100@ohm.york.ac.uk

**ABSTRACT**

A novel method is presented for modifying the spatial information contained in the output from a stereo coincident pair of microphones. The purpose of this method is to provide additional decorrelation of the audio at the left and right replay channels for sound arriving at the sides of a coincident pair but to retain the imaging accuracy for sounds arriving to the front or rear or where the entire soundfield is highly correlated. Details of how this is achieved are given and results for different types of soundfield are presented.

## 1. INTRODUCTION

There are two established techniques for using microphones to capture spatial information for presentation via two-loudspeaker stereo. The first uses two microphones which are coincident in space (or as near is practicable) and which are directional. With this approach the spatial position of sound sources is encoded in the level differences (LD) between the two microphones due to their different relative sensitivities to sound waves arriving from different directions. The second approach uses microphones which are spatially separated ('spaced'), often (although not always) with omnidirectional responses. Here the spatial information is encoded as relative differences in the time of arrival (TOA) of sound waves at the microphones and also, to a limited extent depending on the proximity of the sources to the microphone array, as LD as a result of acoustic attenuation due to the inverse square law. Where directional microphones are used there will be additional differences in directional sensitivity. The 'near-coincident' family of techniques can be seen as a combination of these approaches where spatial information is captured as both LD and TOA. Pure LD techniques give excellent imaging quality when reproduced over loudspeakers yet some listeners report a preference for the 'spatial impression' which is achieved with the TOA approach [1].

There has been recent interest in adaptive processing techniques which can be applied to coincident or ambisonic recordings in order to adapt the delivery of the audio according to how spatially diffuse it is (e.g. [2], [3]). The method described in this paper is designed for use with the Blumlein pair (dipoles at 90 degrees to

each other). This is perhaps not an obvious choice of microphone array for this kind of processing since this particular arrangement has a theoretical correlation coefficient of 0 in a perfectly diffuse field [4]. However reasons why this kind of processing might be appropriate for this kind of array are outlined in the next section, along with an overview of the aims of the proposed algorithm and its implementation. Section 3 describes in detail the novel spatial reconfiguration algorithm that has been developed. Section 4 presents results for various types of signal, both synthetic and acoustic. The final section summarises the paper and presents conclusions and areas for future work.

## 2. SPATIAL ENCODING BY TWO-MICROPHONE ARRAYS FOR STEREO

### 2.1. Overview of array types

Stereo audio as a consumer format has been widely available since the introduction of 45/45 cut vinyl disks in the late 1950s, anticipated by Alan Blumlein's pioneering work twenty years earlier [5]. Whilst surround audio for more than two loudspeakers is becoming more prevalent, the two physical formats designed to deliver audio-only surround (DVD-Audio and SACD) have not seen widespread adoption and it seems that two loudspeaker stereo will be the most common domestic listening format for the foreseeable future [6].

In his landmark patent Blumlein recognised that to produce an accurate 'phantom' image of a point source between two loudspeakers, where that phantom image corresponds to the position of the actual source in the front quadrant of the microphone array, level differences should be captured. Upon replay over loudspeakers these level differences are converted into timing differences between the ears which correspond to those which would have been produced by the actual acoustic source in the same position. This is because, considering a sinusoidal decomposition of sound, the relative phase shift produced at the ears by level differences between two loudspeakers is the same as that produced by path length differences from a single point source at the location of the phantom image. These phase differences are the cues used by the auditory system for localisation below about 1.5 kHz.

Whilst this was the approach adopted for stereo recording by EMI, for whom Blumlein worked, at Decca an adaptation of their method for recording monophony, The Decca Tree, was devised and employed [7]. This is a spaced configuration of omnidirectional microphones which, whilst not offering the point-source imaging accuracy of coincident pairs when replayed over loudspeakers (since the relative delays at the ears become frequency dependent), has been preferred by some engineers and listeners for the sound quality of the recorded sources and for spatial impression. Compromises between coincident and spaced configurations include the NOS (Holland Radio, two cardioid microphones at 90 degrees, 300mm spacing) and ORTF (French Radio, two cardioids at 110 degrees, 170mm spacing) pairs. Because the original Blumlein pair picks up energy equally from all directions and offers excellent imaging of point sources it has earned a reputation as a (or even *the*) purist technique [8]. Despite this, when surveying the evolution and the current state of microphone arrays for two-channel stereo it becomes apparent that no one array is best suited to all acoustic sources, in all buildings, for every listener. Therefore, whatever the microphone array there is a potential benefit from systems that are capable of reconfiguring or transforming the presentation of spatial information in some way. That is the motivation for the work described in this paper which is part of a larger research program concerned with different spatial transformations of audio captured using various kinds of microphone array.

### 2.2. Physics and perception of room acoustics

The response of a room to a sound within it depends on its construction, geometry and ambient conditions. All non-anechoic room responses will consist of a temporally sparse set of early reflections followed by a denser set of secondary, tertiary etc. reflections which becomes progressively denser. The specularity of the reflections and their temporal, frequency and spatial distribution are determined by room and it is usually desirable that at least the later part of that response is largely diffuse in these three domains. A perfectly diffuse field is one in which sound is arriving from all directions with equal probability, a consequence of which is a lack of standing waves.

Research has shown that binaural dissimilarity is a factor in listener preference in room acoustics: a room whose response leads to greater dissimilarity at the ears than another room at a typical listener position is more likely to be preferred [10]. Concert halls whose first reflections at the listener are lateral (i.e. from walls) are more likely to possess this attribute of dissimilarity than

halls where the first reflection is from the ceiling (and is therefore more likely to arrive at both ears at the same time). The proposed solution to the problem of undesirable early reflections from low ceilings in this work is the use of diffusers on these surfaces. In further work by one of the authors it is reported that it is the *magnitude* of the similarity that is the important feature of binaural presentation: negative similarity (i.e. where the signals are the same but in opposite phase) also led to a low preference [11]. This suggests that it is the binaural coherence which is the useful objective parameter here, with a low preference for high coherence and vice versa. A useful discussion of coherence versus correlation in this context can be found in [4]. All of this points to an apparent preference for binaural presentation that is non-coherent for all but direct sound.

### 2.3. Spatial encoding of diffusion

The presentation of multi-channel sound has been the subject of much interest, in both practice- and theoretical-based research, throughout the history of reproduced sound [5]. Recently there has been significant interest in the separate treatment of diffuse and non-diffuse parts of the soundfield.

In [2] ambisonic signals are used to provide estimates of the ratio of acoustic intensity to the total soundfield energy. Since intensity is a vector quantity which represents the net power flowing through a unit area in a particular direction, this ratio provides a measure of the diffuseness (or specularity) of the acoustic waves travelling though a point in space measured by an ambisonic microphone. Where the ratio is 1 all of the soundfield energy is associated with a single plane wave travelling in one direction. Where it is 0 there is no net flow of energy, indicating that the soundfield is diffuse. This ratio is calculated independently for every point on a time frequency grid and used to control how audio is delivered to loudspeakers. The proportion of the energy which is due to directional sound is distributed to loudspeakers in phase (i.e. with no time differences between them) with the direction encoded as level differences between them. The rest of energy is distributed as diffuse sound – each speaker receives the same energy but the phase is randomised so that the presentation at each is decorrelated from that at all of the others. This combines the imaging accuracy of pure level difference stereo with the 'envelopment' by the diffuse sound that is often more commonly associated with time difference panning methods.

In [3] the diffuseness of a soundfield captured by two coincident cardioid microphones is estimated by measuring the normalised cross-correlation coefficient between the outputs from the two microphones at points on a time-frequency grid. This measure of diffuseness is then used to vary the effective directivity patterns of the microphones. This is achieved by the (partial or complete) removal of parts of the signal at the output of one microphone that are also present at the output of the other. The intended effect is to increase the effective directivity of the microphones for non-diffuse sound and to increase the rear pick-up for direct sound, which the authors report as improving presentation. Subjective testing supports this however the tests were conducted via headphone reproduction, rather than loudspeaker reproduction with which coincident microphone techniques are intended to work. (Headphone reproduction of coincident microphone recordings does not produce images outside of the head due to the lack of crosstalk between the ears of the left and right signals.)

Both of these approaches to the presentation of spatial audio are based on the notion that the optimal delivery method for spatial information depends upon the *type* of soundfield being reproduced. These are not the first examples of content-dependent delivery. For example, some Dolby technologies have used this as part of strategies for multi-channel lossy compression, and multi-mic'ed/multitrack recordings with post processing offer all manner of combinations of time-based and level-based presentation. However these more recent approaches are part of a new generation of technologies aimed at presentation enhancement, rather than compression, of existing coincident microphone signals.

### 2.4. Post-processing of audio from a Blumlein pair

As already stated, the advantage of the Blumlein pair is that it does not favour any direction in terms of its energy sensitivity and it provides accurate imaging for the front quadrant, with no bunching of sources either in the centre or at each of the speakers, if those sources are spread evenly across the quadrant. The omnidirectional energy sensitivity of the pair leads to a satisfying presentation of reverberation with a clarity of instrumental line which has been claimed over 'multi-mic'ing plus reverberation' techniques [9]. That said, there are some potential disadvantages. If the response of the room is not satisfying (i.e. it is overly reverberant or has a high critical frequency and so is dominated by strong individual room modes) then the array will not

favour the source over the environment as the more directional response of, for example, a pair of cardioids at 90 degrees ($\pi/2$ radians) would do. A Blumlein pair is shown in Figure 1.

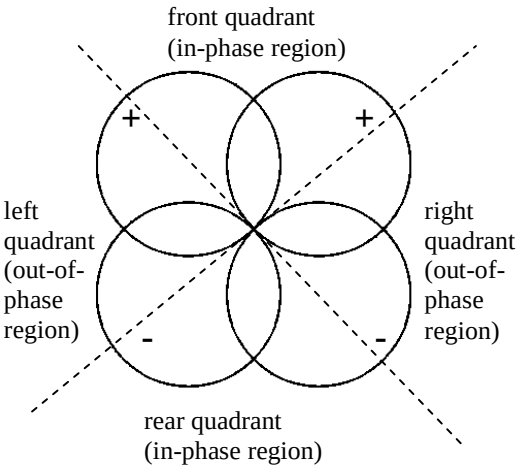



Figure 1: A Blumlein pair (coincident figure-of-eight microphones at 90 degrees to each other). The +/- signs indicate the polarity of the output for positive pressure change approaching each microphone lobe).

Sound sources appearing at the rear quadrant are presented in the front quadrant but are left-right reversed. Sounds arriving at the side quadrants are presented out-of-phase which leads to ambiguous imaging, which may be of benefit if that sound is mainly diffuse reverberation but not if it is an early reflection. As we move a single point source around the quadrants clockwise, starting with the front, we move have an interchannel correlation of 1 (and an interchannel coherence of 1), then a correlation of -1 (coherence of 1), then a correlation of 1 (coherence 1) and finally a correlation of -1. For a perfectly diffuse soundfield (i.e. sound arriving with equal probability from all directions) the total correlation is 0. This is intuitively explained in [12] as the result of half of the soundfield arriving at the microphone pair having a correlation of 1 and the other half having a correlation of -1, the addition of which gives an overall correlation of 0.

For a point source arriving from directly in front of the pair the amplitude and phase at both microphones are identical. If that source moves so that it is at an angle of 45 degrees ($\pi/4$ radians) to the median plane then it is on the front axis of one microphone and on the null axis of the other, so there is maximum output from one microphone and zero output from the other. As the source moves further to the side there is output from both microphones again, but this time they are in anti-phase. The representation of 'fully to the side' (i.e. at 90 degrees to the median plane) is for both microphones to have the same output amplitude but opposite phase. The fact that the left and right quadrants are out-of-phase is considered advantageous in many recording situations, since that sound arriving in these quadrants is indirect (reflected/reverberant) sound, with direct sound from performers arriving in the front quadrant and that from the audience (applause etc.) arriving in the rear. Since the side quadrants are out of phase this can give rise to reverberation that spreads outside of the speakers, giving rise to a greater sense of envelopment or spatial impression, in a similar way to that which occurs with spaced configurations, but cannot occur with 'always in-phase' configurations such as those which only employ cardioids. However it should be remembered that for sounds arriving directly to the side, coincident figure-of-eights, unlike spaced configurations, are presenting the same waveforms at the same time, just in opposite phase. Considering the preference for a correlation of 0 over 1 *or* -1 for lateral reflections in a concert hall discussed in Section 2.2, this suggests that temporal separation of discrete side reflections, as offered by spaced pairs, would be preferred over the out-of-phase but temporally coincident presentation by Blumlein pairs. At this stage it has to be acknowledged that preferences for microphone arrays are a subjective matter. For this author the presentation of a sound at 90 degrees to the median plane in an anechoic environment by a spaced pair of omnidirectional microphones is preferable to that offered by a Blumlein pair. The latter provides an image which is imprecise and can appear to move from side to side, whereas the spaced pair presentation gives an image which, whilst not heard at 90 degrees to the right of the listener, does not meander.

The motivation for the work described in this paper is a desire to combine the excellent imaging quality and equal energy capture of the Blumlein pair with the preferred presentation for sounds from the side and sense of envelopment often associated with spaced pairs. This has been attempted not through the development of a new microphone configuration, but via an algorithm which is able to modify the representation of spatial information within an existing Blumlein pair recording. It is inspired by a personal preference as a listener and recording engineer rather than a conviction of what is definitely and objectively

the best approach to spatial presentation of two channel audio via two loudspeakers.

## 3. ALGORITHM FOR ADAPTING SPATIAL INFORMATION FROM BLUMLEIN PAIRS

The algorithm described in this paper aims to offer an alternative presentation of spatial information captured by Blumlein pairs. An overview of the method is:

1. Decompose the signal from both microphones into a time-frequency representation via the short-time Fourier transform (STFT).
2. Locate the dominant component in the front/rear quadrant (figure-of-eight pairs are unable to differentiate between opposing quadrants).
3. Rotate the pair so that the middle points towards this component. Considering the equivalent mid-side version of the pair, this means that the source is pointed to directly by the M microphone and the null axis of the S microphone.
4. Determine the similarity between the M and S signals by correlation analysis.
5. Where the signals are perfectly (negatively or positively) correlated then the S signal should also be panned to its correct position by level difference panning (this is equivalent to performing no processing of the signals at all).
6. Where there is no similarity, pan the M signal to the correct position between the speakers by purely level difference panning, deliver the S signal with equal level to each loudspeaker but with temporal separation.
6. Where the magnitude of the correlation is between 0 and 1 a combination of 4 and 5 should be applied.

### 3.1. Time-frequency analysis

The short-time Fourier transform (STFT) is used to perform time-frequency analysis with frame lengths being either 1024 or 2048 samples with a sampling rate of 44.1 kHz. A Hann window is used. Zero-padding is also employed to over-sample the spectrum which gives a finer frequency grid (although, of course, it is the frame length and the window function that determine the actual frequency resolution). Also, the use of zero-padding is essential if linear, rather than circular, time-shifts are to be introduced into the signals in the Fourier domain, as is the case with the algorithm described here. A zero-padding factor of 8 is employed here and the overlap factor is 2 (e.g. for a frame length of 1024 the hop size is 512).

Although the STFT produces a constant bandwidth analysis, individual frequency bins are combined into groups which correspond to the equivalent rectangular bandwidth (ERB) of the auditory filter. The ERB scale is defined by

$$\mathrm{ERB}=21.4\log_{10}\left(0.00437\,f+1\right) \tag{1}$$

where ERB is the auditory band into which the frequency $f$ falls [13]. Rearranging this equation the lower band edges for the $n$th ERB, in terms of the nearest Fourier bin, are given by:

$$\mathrm{lower}_n=\mathrm{round}\left(\frac{10^{\frac{n-1}{21.4}}-1}{0.00437}\frac{N}{F_s}\right), n=1,2,3... \tag{2}$$

where $N$ is the zero-padded size of the analysis frame, $F_s$ is the sample rate and $n$ runs until the Nyquist limit is exceeded. A benefit of using a high zero-padding factor is that the position of the band edges will correspond more closely to the position of the ERB band edges, since the resolution of frequency axis of the Fourier analysis is finer.

### 3.2. Dominant component analysis

Having divided the Fourier spectrum for a single frame into groups of bins that correspond to ERBs, the algorithm then proceeds to identify the direction of the dominant component in each ERB. The first step in this process is transforming the left and right signals of the microphone pair into the equivalent middle (M) and side (S) signals (i.e. those that would have been generated by an M microphone pointing directly forwards and an S signal at 90 degrees to that).

$$M_n=\sum_{k=\mathrm{lower}_n}^{\mathrm{lower}_{n+1}-1}\left|R(k)+L(k)\right| \tag{3}$$

$$S_n=\sum_{k=\mathrm{lower}_n}^{\mathrm{lower}_{n+1}-1}\left|R(k)-L(k)\right| \tag{4}$$

where $R(k)$ and $L(k)$ are the $k$th bins of the Fourier transforms of the right and left microphone signals respectively. (Of course, these M and S signals could be derived in the time domain prior to Fourier analysis.) Although there may be many sources within a single ERB contributing to the energy within that band, and

they may be at different positions, the simplified interpretation of the data is that there is a single dominant component within a single ERB in a given analysis frame. The angular direction, $\theta$, of that dominant component is given by:

$$\theta = \arctan\left(\frac{S_n}{M_n}\right) \quad (5)$$

Since S and M are calculated using absolute values the sign of $\theta$ must be determined by comparing the amount of energy in the left channel with that in the right: where there is more energy in the right channel $\theta$ is positive and where there is more energy in the left channel it is negative. As stated before, for purely figure-of-eight pairs there is ambiguity between the front and rear and left and right quadrants. Therefore the range of $\theta$ is +/- $\pi/2$ radians. Additionally this algorithm constrains $\theta$ to be in the range +/- $\pi/4$ radians so that a dominant component can only exist in the front\rear quadrants. The motivation for this constraint will be described in the next sub-section.

Having determined the direction of the dominant component within an ERB group, the next step is to derive a new pair of signals which correspond to a figure-of-eight microphone pointing in the exact direction of the dominant source, and an accompanying microphone at 90 degrees to this. This second microphone will have its null axis pointing in the direction of the dominant source. If the dominant source is indeed a single source, and it behaves as a point source then no direct sound from that source will be picked up by the second microphone. If there is a signal at the second microphone then this is an indication that energy for that component is not arriving from a single direction and therefore that it is, to a certain extent, diffuse. These virtual microphone signals are derived using:

$$M_{\text{dominant},n}(k_n) = \frac{1}{\sqrt{2}}\left(\cos\theta - \sin\theta\right)\left(R(k) + L(k)\right) \quad (6)$$

$$S_{\text{dominant},n}(k_n) = \frac{1}{\sqrt{2}}\left(\cos\theta + \sin\theta\right)\left(L(k) - R(k)\right) \quad (7)$$

where $n$ indicates the ERB and $k_n$ is an integer that runs from 0 to ($\text{upper}_n$ – $\text{lower}_n$) and is the index of the Fourier bins within the $n$th ERB.

### 3.3. Spatial reconfiguration

Having identified the dominant component direction for every ERB within a frame and derived new signals which correspond to an M-S pair pointing at this component, the next stage of the process is to adaptively adjust the presentation of the spatial information captured in these two signals. The aims of the algorithm are to:

1. Preserve the level-difference presentation of components that appear in the front/rear quadrant since this is where direct sound from performers will arrive from.

2. Present sounds arriving from the side quadrants using time differences in order to provide decorrelation between left and right channels for diffuse (and near-diffuse) sound and to avoid 'time-coincident but out-of-phase' presentation of lateral reflections (as discussed in sub-section 2.4).

These aims are achieved by:

1. Level difference panning (using the cosine law) of the $M_{\text{dominant}}$ signal.

2. Adaptively introducing time difference panning to the $S_{\text{dominant}}$ signal. The proportion of $S_{\text{dominant}}$ for a particular ERB that is time difference panned, as opposed to level difference panned, is determined by the absolute normalised spectral cross-correlation, $R$, between $M_{\text{dominant}}$ and $S_{\text{dominant}}$ calculated across the Fourier bins that fall within that ERB:

$$R_n = \frac{\left|\sum_{k=\text{lower}_n}^{\text{lower}_{n+1}-1} M_{\text{dominant},n}(k_n)\, S*_{\text{dominant},n}(k_n)\right|}{\sqrt{\sum_{k=\text{lower}_n}^{\text{lower}_{n+1}-1}\left|M_{\text{dominant},n}(k_n)\right|^2 \sum_{k=\text{lower}_n}^{\text{lower}_{n+1}-1}\left|S_{\text{dominant},n}(k_n)\right|^2}} \quad (8)$$

This gives a value of $R_n$ in the range 0 to 1. The $S_{\text{dominant}}$ signal is then distributed to the two panning methods according to the following ratios:

$$S_{\text{dominant, } n\text{, level-difference}} = S_{\text{dominant},n} R_n^p \quad (9)$$

$$S_{\text{dominant, } n\text{, time-difference}} = S_{\text{dominant},n}\left(1 - R_n^p\right) \quad (10)$$

where $p$ is a parameter which can be used to control the amount of spatial reconfiguration, although in practice a default setting of $p$ = 1 has found to be generally satisfactory.

The purpose of adaptively assigning $S_{\text{dominant},n}$ to a panning method is to avoid audible artifacts where there is more than one frontal source and the ERB spectra of these sources overlap. The worst-case scenario here is that of two sources being at angles of +/- 45 degrees of the forward direction of the pair (i.e. at the extreme edges of the front quadrant). In this case both $M_{\text{dominant},n}$ and $S_{\text{dominant},n}$ will contribute to the signals representing each front source and here there should be no time-difference panning, since this will lead to audible artifacts due to the temporal spreading of these direct sounds. To avoid this, where there is correlation between the two signals within an ERB, then the amount of time-difference panning of $S_{\text{dominant},n}$ should be reduced. In the limit, where $R_n$ = 1, there is no spatial reconfiguration and the output of the algorithm is the same as the input (albeit with a short delay due to the Fourier analysis and resynthesis). The reasoning behind the choice of spectral cross correlation rather than a coherence measure is that $R_n$ will be close to 1 where the $M_{\text{dominant},n}$ and $S_{\text{dominant},n}$ have the same magnitude profiles and are in either in-phase or out-of-phase. Where they have different magnitude profiles or their phase difference is closer to 90 degrees (which would not be the case for a non-diffuse soundfield) $R_n$ will be closer to 0.

### 3.4. Panning and signal reconstruction

The signals are recombined in the Fourier domain according to:

$$L_{\text{processed},n} = \frac{1}{\sqrt{2}} \begin{pmatrix} \cos\theta \left( M_{\text{dominant},n} + S_{\text{dominant},n,\text{level-difference}} \right) \\ +\sin\theta \left( S_{\text{dominant},n,\text{level-difference}} - M_{\text{dominant},n} \right) \\ +S_{\text{dominant},n,\text{time-difference}} z^{-\tau_L} \end{pmatrix} \quad (11)$$

$$R_{\text{processed},n} = \frac{1}{\sqrt{2}} \begin{pmatrix} \cos\theta \left( M_{\text{dominant},n} - S_{\text{dominant},n,\text{level-difference}} \right) \\ +\sin\theta \left( M_{\text{dominant},n} + S_{\text{dominant},n,\text{level-difference}} \right) \\ +S_{\text{dominant},n,\text{time-difference}} z^{-\tau_R} \end{pmatrix} \quad (12)$$

where $z$ is the time shift operator ($z^{-1}$ is equivalent to a delay of 1 sample) and $\tau_L$ and $\tau_R$ are the delays applied to left and right channels respectively to produce the time-difference panning. The $1/\sqrt{2}$ term ensures that there is no change in energy at the output. Where $M_{\text{dominant},n}$ is panned towards the right then the delay applied to $S_{\text{dominant},n,\text{time-difference}}$ in the right channel is more than the delay applied to it in the left channel and vice versa. The reasoning behind this is that, whatever direction $M_{\text{dominant},n}$ is perceived as arriving from, $S_{\text{dominant},n}$, should be perceived as coming from the opposite side of the speaker array (as happens when it is level-difference panned), which means that $S_{\text{dominant},n,\text{time-difference}}$ should lead on that side. In case $S_{\text{dominant},n,\text{time-difference}}$ contains any of $M_{\text{dominant},n}$ the delay applied to $S_{\text{dominant},n,\text{time-difference}}$ relative $M_{\text{dominant},n}$ is always positive so that the risk of pre-echo of the dominant component is minimised.

The delays applied are calculated according to:

$$\tau_L = \left(1 - \cos(\theta)\,\text{sgn}(\theta)\right) F_s D \quad (13)$$

$$\tau_R = \left(1 + \cos(\theta)\,\text{sgn}(\theta)\right) F_s D \quad (14)$$

where $D$ is the mean of the delays applied to both channels for $S_{\text{dominant},n,\text{time-difference}}$ relative to $M_{\text{dominant},n}$ and 2$D$ is the maximum delay that can be applied to a channel. The 2$D$ case occurs when the sound is arriving to the side of the original pair used for the recording, and the distance 2$D$/$c$, where $c$ is the speed of sound in air, is the distance between two spaced microphones that would capture the same relative delay between channels for sound arriving directly to the side of such a pair. This is shown in Figure 2.

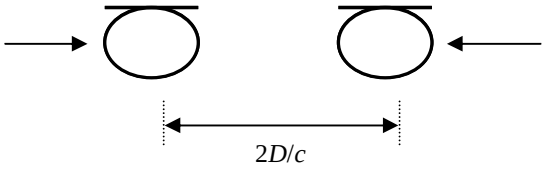



Figure 2: Equivalent microphone spacing for sound arriving from +/- 90 degrees to straight ahead

$\tau_L$ and $\tau_R$ are plotted against $\theta$ in Figure 3. The large discontinuities at $\theta$ = 0 are necessary to meet the criteria described previously for time-difference panning. Firstly, when $M_{\text{dominant},n}$ is pointing directly forwards then $S_{\text{dominant},n}$, is arriving directly from the side, requiring the largest delay difference between the signal that is fed to the left and right channels. Secondly, when

$\theta$ is positive, and $M_{\text{dominant},n}$ level-difference panned towards the right channel, then $S_{\text{dominant},n,\text{time-difference}}$ should lead in the left channel and, when $\theta$ is negative, $S_{\text{dominant},n,\text{time-difference}}$ should lead in the right channel. Both of these requirements combine to give the discontinuity at $\theta = 0$. This is undesirable but is an unavoidable consequence of the bi-polarity of the figure-of-eight response. The value of $D$ can be set according to taste, although a default of 20 ms, which gives a maximum interchannel delay for $S_{\text{dominant},n,\text{time-difference}}$ of 40 ms, has been used when testing the algorithm. Whilst this delay is far too large to represent the physical distance between a main pair of spaced microphones at nearly 13 metres it does feasibly represent the distance between spaced ambience microphones.

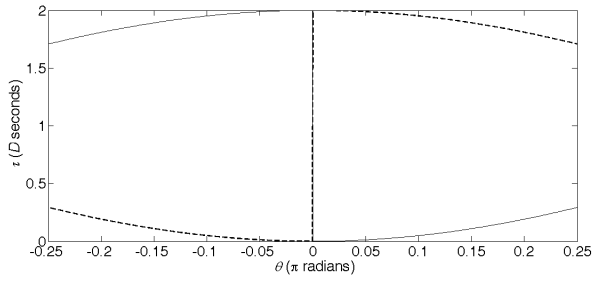



Figure 3: Delays applied to $S_{\text{dominant},n,\text{time-difference}}$ for left channel (solid line) and right channel (dashed line) for angle of incidence ($\theta = 0$ is sound arriving at centre) of $M_{\text{dominant},n}$.

## 4. EVALUATION

In this section the behaviour of the algorithm is presented for some simple synthetic test signals. A subjective description of the processing of acoustic signals using Blumlein pairs is also given. A full evaluation of a process for altering the spatial quality of recorded audio should, of course, include the results of listening tests of distinguishability, preference etc. As stated, this is part of a larger effort which is considering many different approaches to spatial reconfiguration of recordings made using many different microphone arrays. As such, at this stage, large-scale listening tests are deferred until there are more approaches to test (possibly including improvements to the process described in this paper). However, audio examples will be played at the oral presentation of this paper and these can also be found online [14].

### 4.1. Test signals

The purpose of this sub-section is to illustrate that the algorithm functions as described in the Section 3. The test signals are synthetic and chosen for their ability to demonstrate different aspects of the system rather than because they are representative of a particular kind of commonly encountered acoustic signal. For each of the signals in this sub-section the sampling frequency is 44.1 kHz, the analysis frame length is 1024, zero-padded to 8192 samples. The Hann window is used. Figure 4a shows the effect of the algorithm for a mono impulse. As intended there is no change between input and output, except for some noise as a result of the numerical processing (Fourier transform, derivation and rotation of each $M_{\text{dominant},n}$ and $S_{\text{dominant},n}$ followed by re-panning and inverse Fourier transform). This noise is not visible in the plots and peaks at -78 dB below the level of the impulse.

Even if the correlation measurement is bypassed (i.e. $R_n$ is forced to 0) then the output is still that shown in Figure 4a since there is single component (no diffusion of sound at any frequency) and that component arrives in the front quadrant. Figure 4b illustrates what happens when sound arrives in the side quadrants with the correlation measurement. Here the direction of the dominant component is constrained to $\pi/4$ radians and this is represented by the $M_{\text{dominant},n}$ part of the right channel signal (circled in the figure). The $S_{\text{dominant},n}$ component is entirely time-difference panned (since $R_n$ has been forced to 0) and the delays applied to the impulse in the left and right channel correspond equations (13) and (14) respectively (and Figure 3). The spreading of the impulses is a result of non-integer sample shifts. Energy is preserved, the output energy is within 0.02 dB of the input energy.

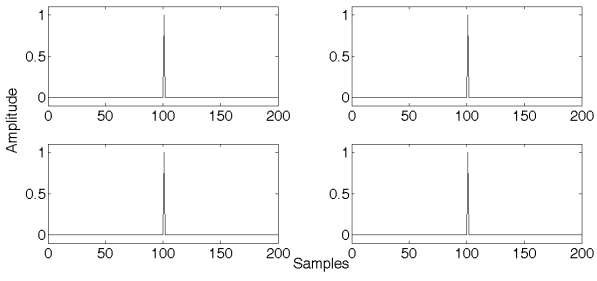



Figure 4a: Input (left panels, left channel is top panel, right channel is bottom panel) and output (right panels, left at top, right at bottom) for an in-phase impulse.

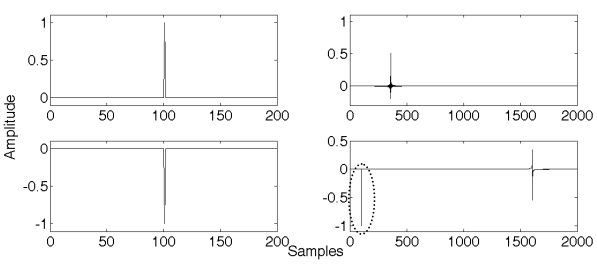
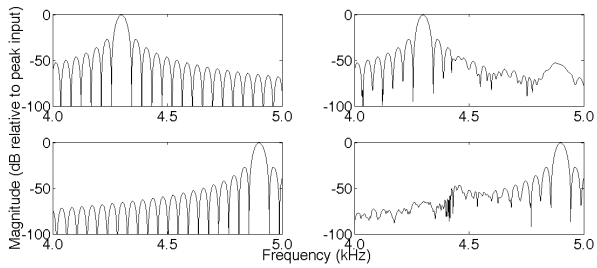



Figure 4b: Input and output (panel layout as for Figure 4a) for an out-of-phase impulse with correlation forced to 0.

In order to show all aspects of the system functioning on a simple test signal (i.e. without $R_n$ being fixed at 0) Figure 5 shows the effect of the algorithm on a combination of two sinusoids. The two sinusoids fall within one ERB (the 29th) but are at different frequencies (4.3 and 4.9 kHz) and appear in different channels (left and right respectively). The magnitude of the Fourier transform of the left (top) and right (bottom) inputs (left panels) to the system are shown in the figure. The right panels show the corresponding outputs where it can be seen that the original sinusoidal components have been preserved but diffuse versions of the component from the opposite channel also now appear. This is due to the dominant direction being straight ahead (as there is equal energy in each channel within the same ERB) but *Rn* is less than 1 since the Fourier spectra within that ERB are not the same.




## 4.2. Processing of recordings of acoustic events

In this section the perceived effect of the algorithm on signals recorded in an actual acoustic space is discussed. In two of these recordings the sound source is electroacoustic, in another the source is acoustic. They were all made by the author at the National Centre for Early Music (NCEM) in York, England [15].

Firstly the recording of the response of the NCEM to an impulse is considered. This was recorded via the sine-sweep method of Farina [16] using AKG Blueline figure-of-eight microphones in a mid-side configuration with the mid microphone pointing directly at the sound source at a distance of about 4.5 m. The difference here is subtle but there is a perceptible increase in the 'separate-ness' of the two loudspeakers upon replay although there is very little change in the ratio of mid to side signal. In fact this ratio is higher for the signal output from the algorithm (7.4 dB) than it is for the input (6.8 dB). There is some just noticeable smearing of the onset of the impulse in the output.

The next example is a recording of a loudspeaker playing an anechoic recording of a guitar in the NCEM. The microphone array is the same as for the first recording but the source is almost fully left within the front quadrant and at slightly reduced distance of 4 m. In this case the effect is difficult to perceive at times with an analysis frame of 1024 samples and a greater sense of separation and envelopment is achieved with a frame size of 2048. This is perhaps because the source is more stationary and so a longer frame length is required to give sufficiently low values of $R_n$ for the effect of the algorithm to be audible. At one or two points in the output a very quiet repeat of the guitar pluck onset can be heard. This is barely audible but does reduce the acoustic plausibility of the output.

The final NCEM recording is of a four part male vocal ensemble. This was recorded in ambisonic B-format using a Soundfield microphone and subsequently processed to derive a Blumlein pair recording. The ensemble is equally across an arc which spans about two thirds of the front quadrant. The singers are each about 4 m away from the microphone. The brief excerpt chosen is of continuous singing with either three or all four parts present. Here there is little perceptible difference between input and output at 1024 or 2048, although at 2048 an undesirable shimmering effect is audible during a sustained chord, which also occurs when $p$ in equation (10) is increased from 1 to 3.

These examples demonstrate that the algorithm described in this paper is successful in changing the sense of spaciousness due to reverberation in acoustic recordings, however there are some just audible artifacts.

## 5. CONCLUSIONS

This paper has described an algorithm for changing the presentation of space in recordings made with a Blumlein pair. The algorithm works by identifying and rotating to the dominant direction of sound at points on a time-frequency grid. The frequency resolution of the grid is related to the critical bands of the human ear although this is derived from an STFT. The dominant source direction is constrained to be within the front/rear quadrants of the microphone pair. The mid signal obtained when the array is rotated to the dominant direction is level-difference panned. The side signal obtained after rotation is panned via a combination of time-difference and level-difference panning – the ratio in which these two panning operations are combined is determined by the spectral cross-correlation of the signal at that point in time-frequency.

Informal listening has demonstrated that there are audible differences in the presentation of reverberation when the process is applied to existing Blumlein pair recordings although there can be some minor artifacts. Further work will focus on eliminating these artifacts. Possible solutions include post-processing of the output signal so that variations in the output energy in each ERB better match that at the input or adaptive variation of *D* or the analysis frame length.

The intention is not to 'improve' the quality of presentation of audio recorded using a Blumlein pair, but to offer an alternative presentation of un-correlated sound arriving at the side quadrants so that there is a greater sense of separation between the two loudspeakers. Of course, for many listeners the spatial information captured by this kind of coincident pair is already optimal. Whether the differences in presentation offered by the processing described in this paper are desirable is left to the ears of the beholder.

## 6. REFERENCES

[1] Rumsey, F., "Subjective Assessment of the Spatial Attributes of Reproduced Sound", *Paper presented at the Audio Engineering Society 15th International Conference*, 1998.

[2] Pulkki, V., "Spatial Sound Reproduction with Directional Audio Coding", *Journal of the Audio Engineering Society*, vol. 55, 2007, pp. 503-516.

[3] Faller, C., "Modifying the Directional Responses of a Coincident Pair of Microphones by Postprocessing", *Journal of the Audio Engineering Society*, vol. 56, 2008, pp. 810-822.

[4] Martin, G., "A New Microphone Technique for Five-Channel Recording", *Paper presented at the Audio Engineering Society 118th International Convention*, 2005.

[5] Torick, E., "Highlights in the History of Multichannel Sound", *Journal of the Audio Engineering Society*, vol. 46, 1998, pp.27-31.

[6] Recording Industry Association of America, "2008 Year-end Shipment Statistics", available at http://www.riaa.com/keystatistics.php

[7] Gray, M., "The Birth of Decca Stereo", *The Journal of The Association of Recorded Sound Collections*, 1986, pp. 4-15.

[8] Lipshitz, S., "Stereo Microphone Techniques: Are The Purists Wrong?", *The Journal of the Audio Engineering Society*, 1986, pp. 716-744.

[9] Gerzon, M., "Stabilising Stereo Images", *Studio Sound*, December 1974.

[10] Schroeder, M., Gottlob, D. and Siebrasse, K.F., "Comparative Study of European Concert Halls: Correlation of Subjective Preference with Geometric and Acoustic Parameters", *Journal of The Acoustical Society of America*, October 1974, pp. 1195-1201.

[11] Schroeder, M., "Binaural Dissimilarity and Optimum Ceilings for Concert Halls", *Journal of The Acoustical Society of America*, April 1979, pp. 958-963.

[12] Martin, G., *Introduction to Sound Recording*, available at www.tonmeister.ca, 2006.

[13] Moore, B. and Glasberg, B., "A Revision of Zwicker's Loudness Model", *Acta Acustica*, vol. 82, 1996, pp. 335-345.

[14] Audio examples for this paper available at: www.jezwells.org/spatial_reconfiguration

[15] National Centre for Early Music website: www.ncem.co.uk

[16] Farina, A., "Simultaneous Measurement of Impulse Response and Distortion With a Swept-Sine Technique", *Paper presented at the Audio Engineering Society 108th International Convention*, 2000.