

Localization Accuracy of Advanced Spatialization Techniques in Small Concert Halls

Enda Bates,^{a)} Gavin Kearney,^{b)} Frank Boland,^{c)} and Dermot Furlong^{d)}
Department of Electronic & Electrical Engineering, Trinity College Dublin, Ireland

(Dated: May 30, 2007)

A comparison of spatialization schemes is presented in terms of their localization accuracy under the non-ideal listening conditions found in small concert halls. Of interest is the effect of real reverberant conditions, non-central listening positions and non-circular speaker arrays on source localization. The data is presented by comparison of empirical binaural measurements to perceptual listening tests carried out using Ambisonics, Vector Base Amplitude Panning (VBAP), Spat (with B-format encoding) and Delta Stereophony (DSS) systems. The listening tests are conducted by comparing the localization of phantom sources generated by the spatialization systems, to monophonic sources generated by reference loudspeakers. The reference and phantom sources are presented at front, side and back locations about a 9 listener audience, and the systems are tested in a random order with a calibrated 16 loudspeaker array situated around the audience area. The binaural recordings are compared to the subjective measurements of localization accuracy through the inter-aural time difference (ITD) cues at each listener position.

Keywords: Sound localization, spatialization, distributed audience, Ambisonics, VBAP, Spat, Delta Stereo

I. INTRODUCTION

There are numerous difficulties associated with the presentation of spatialized audio to a distributed audience in a concert hall environment. The spatial locations and trajectories created by the composer in the studio may not be experienced by the eventual audience due to the non-ideal conditions found in typical concert halls. In particular, the presence of room reverberation and reflections can impact negatively on source localization and on the performance of the implemented spatialization technique. Many of these techniques are also optimized for a single, centrally positioned listener and not for the large listening area required for a concert performance. Finally, the regular symmetrical loudspeaker arrangements recommended for most spatialization techniques can be difficult to implement in a standard rectangular hall without greatly reducing the size of the audience area.

In this paper, we present the results of a comparative study on the localization performance of various advanced spatialization techniques for a distributed audience in a small concert hall. A subjective assessment of source localization accuracy is implemented through a series of perceptual listening tests using an asymmetrical sixteen-channel loudspeaker array. These tests are then compared to binaural recordings taken in the audience area, where virtual source ITDs (Interaural Time Differences) are assessed against monophonic source measurements. We will begin this study with an overview of human auditory localization in reverberant environments followed by a brief summary of the assessed spatialization

techniques.

A. Auditory Localization

Human auditory localization can be broken down into three main categories, namely directional hearing in the horizontal plane, directional hearing in the vertical plane, and “distance hearing”¹. Here we will limit our discussion to the horizontal plane, as this is particularly relevant for most spatialized audio presentations. Two main cues dominate localization in the horizontal plane, namely the ILD (interaural level difference), which arises from the shadowing effect of the head, and ITD, which arises from the spatial separation of the two ears². The variations in localization accuracy for different source stimuli in the free field has been well documented by Blauert, Mac Pherson, Weinrich and Sandel^{1,3-5}. It has been shown that there is a strong weighting for ITDs with low frequency signals and poor weighting of ITDs with high frequency signals. The converse is true for the ILD. For wideband stimuli, the ITD is found to dominate³.

Numerous studies have demonstrated that the region of most precise spatial hearing lies in the forward direction with frontal hearing having an accuracy of between 4.4° and 10° for different signal types¹. Localization ability decreases as the source azimuth moves to the sides, with the localization blur at $\pm 90^\circ$ being between three to ten times its value for the forward direction. For sources to the rear of the listener, localization blur improves somewhat but is still approximately twice that for frontal sources. Our localization accuracy is also dependent on the nature of the source signal and the acoustical environment. Hartmann et al⁶ demonstrated that sources containing strong transients are localized independently of the room reverberation time, but may depend on the room geometry. Conversely, for sounds without attack transients, localization improves monotonically with the

^{a)}Electronic address: batesja@tcd.ie

^{b)}Electronic address: gpkearney@ee.tcd.ie

^{c)}Electronic address: fboland@tcd.ie

^{d)}Electronic address: dermot.furlong@tcd.ie

spectral density of the source. However, localization of *continuous* broadband noise is dependent on room reverberation time. The presence of early reflections has been shown to affect localization accuracy, even for transient signals⁷. Specifically, the early reflections from side walls impact negatively on horizontal localization, while early reflections from the floor and ceiling help to reinforce horizontal localization. It should be noted that this is the opposite of the preferred arrangement for acoustic presentations in concert halls, which emphasizes lateral reflections.

B. Spatialization Methods

In the latter half of the last century, the basic stereophonic principles used in 2 and 3 channel stereophony were extended to create sound systems that surround entire audience areas. Since then, various multichannel systems have provided a medium for engineers and composers to place phantom auditory sources 360° about an audience to enhance program material. In this paper, we assess the localization performance of Second Order Ambisonics, VBAP, Delta Stereophony and Spat (with B-Format Ambisonics decoding).

1. Second Order Ambisonics

Ambisonics was developed by Michael Gerzon as a complete approach to recording, manipulating and synthesizing artificial sound fields for surround and periphonic loudspeaker arrangements^{8,9}. It has been regularly used in spatial music and theatre for the past three decades and an excellent overview of the system can be found in¹⁰. One of the most attractive features of Ambisonics is the separation of the encoding and decoding functions, which means that spatial information can be created and encoded independently of the reproduction setup.

Although Ambisonics was originally developed as a point source solution for a single listener, its recent extension to higher orders of spherical harmonics has suggested that the listening area can be extended significantly. Although very little research has been published on the performance of these systems under concert hall conditions, several Ambisonic decoding schemes and loudspeaker configurations were assessed under anechoic conditions by Benjamin et al¹¹. One interesting conclusion drawn from this study was that changes in loudspeaker layout are significantly more important than changes in the decoding scheme. For the tests presented in this paper, the Second Order Ambisonics encoding and decoding was carried out using the set of externals for Max MSP developed by the Institute for Computer Music and Sound Technology (ICST)¹². These externals were based on a Csound implementation of Ambisonics created by David Malham of York University, who also published one of the few papers on large area Ambisonics systems¹³.

When implementing Ambisonics systems, the directional response pattern of the generated soundfield can be

narrowed or widened according to the needs of the room acoustics, the position of the speakers and required listening area. This is achieved by weighting the various order components differently during the decoding process. Prior to the formal listening tests the authors experimented with different weightings, using the three schemes listed below in Table I as a starting point. The Furse-Malham "matched" weighting increases the directional response which is optimal for a single listener but produced poor results in the test environment. Controlled opposite, or In-Phase decoding reduces the directional response and produces a more diffuse soundfield. Although this is often recommended for large listening areas, this weighting produced very poorly localized sources. Overall, the best localization was achieved using a ratio (recommended by Malham for 8-speaker setups) that lies between "matched" and "controlled opposites" decoding (Entry three in Table I).

Weighting Scheme	Furse-Malham Set	In-Phase	Malham
0 Order	0.707107	1.944	0.823242
1 st Order	1	1.296	1
2 nd Order	1	0.324	0.442259

TABLE I. Ambisonic Decoder Weighting Schemes

2. VBAP

Vector Base Amplitude Panning (VBAP) is a generic method for virtual source positioning developed by Ville Pulkki¹⁴. This vector-based reformulation of the amplitude panning method can be used to extend the basic stereophonic principle to an arbitrary number of loudspeakers. Once the available loudspeaker layout is defined, virtual sources are positioned by simply specifying the source azimuth. If a virtual source is panned to the same direction as any of the loudspeakers, then only that loudspeaker will be used. If a source is panned to a point between two loudspeakers then only those two loudspeakers will be used to produce the virtual source using the tangent panning law. The flexibility of VBAP gives it a distinct advantage over other amplitude-panning based spatialization schemes and its straightforward implementation and scalability make it an attractive solution for spatialization.

3. Spat

Spat is a real-time modular spatial-sound-processing software system developed by IRCAM and Espaces Nouveaux for the Max MSP environment¹⁵. The system allows for the positioning and reverberation of audio sources in three dimensions using a high level control

interface based on a number of perceptual parameters. The design of Spat is largely based on the spatial processing algorithms developed by Chowning and Moore in the seventies and eighties^{16,17}. The supplied output module can be configured for reproduction over loudspeakers using standard stereophony, discrete intensity panning over various multichannel loudspeaker configurations, Ambisonics B-format encoding or binaural encoding for reproduction over headphones. This approach is conceptually similar to that of Ambisonics systems which also separate the encoding and decoding functions. The user can therefore specify the required spatial locations, distances and trajectories using high level perceptual controls and then independently select the output module most suitable for the reproduction environment. In this case, the Ambisonics B-Format scheme was chosen to position the direct sound at the required locations around the audience area, while the level of the artificial reverberation generated by Spat was minimised. It should be noted that, to the authors knowledge, IRCAM have not released any details of the precise Ambisonics encoding and decoding schemes implemented in Spat.

4. Delta Stereophony

Delta Stereophony is a sound reinforcement system intended to provide correct localization of sound sources for a distributed audience. DSS is largely based on the precedence effect and ensures that each listener in an auditorium receives the direct sound from the original sound direction first, before that of reinforcement speakers placed around the audience area¹⁸. In effect this can be considered an analogous to Chowning’s model of spatialization using direct and reverberant sound, with the emphasis being on achieving correct localization for a distributed audience rather than modelling room reverberation. It is the opinion of the authors that this approach is highly applicable for spatial audio presentations in listening environments that already possess significant reverberation characteristics. In addition the DSS system aims to provide uniform sound reinforcement levels about the audience area. The advantage of this scheme over other sound reinforcement systems is that DSS ensures that the delayed loudspeakers do not exceed the upper limits of the precedence effect causing echo suppression of the direct sound. However, the utilization of the reinforcement speakers ensures for a more uniform SPL about the audience area. It is important to note that DSS generally uses monophonic source simulation radiators to reinforce the direct sound¹⁸, ensuring good localization. However, as was shown by Ahnert,¹⁹ it has on occasion been used for the reinforcement of moving sources. In order to compare DSS as a spatialization scheme then, we assess here its ability to form phantom images between source radiators.

II. ASSESSMENT METHODOLOGY

Both subjective and objective methodologies are required in order to fully assess the performance of spatialization techniques in terms of their localization accuracy. Perceptual listening tests are required to determine the localization accuracy of sources positioned at different locations around the test audience. In addition, binaural recordings are useful in assessing the non-perceptual effects of the room acoustics on the performance of the spatialization system. Various source stimuli are used in order to determine the perceptual effect of the spectral and temporal content of the source signal on localization accuracy. Finally, monophonic sources positioned at the same locations as the virtual sources are presented during the test in order to determine the localization accuracy for a *real* source under similar conditions.

A. Test Setup

A small sized concert hall, located in Trinity College Dublin, was chosen as the test room and is shown in Figure 1. The spatially averaged reverberation time (RT60) of the hall was measured using maximum length sequence noise and the range of values over the frequency spectrum is shown in Figure 2. A loudspeaker array consisting of



FIG. 1. Printing House Hall in Trinity College Dublin showing listener/loudspeaker setup.

16 Genelec 1029A loudspeakers was arranged around a 9 listener audience area as shown in Figure 3. A PC utilising a MOTU896 audio interface was used to route the audio to the loudspeakers.

In these tests, monophonic sources were presented using the 8 black loudspeakers shown in Figure 3 while the 8 white loudspeakers were used by the various spatialization techniques to generate virtual sources at the same positions. This method allows for a direct comparison between the localization accuracy for a real source, and a virtual source positioned at the same location. The presented source stimuli consisted of 1 second unfiltered

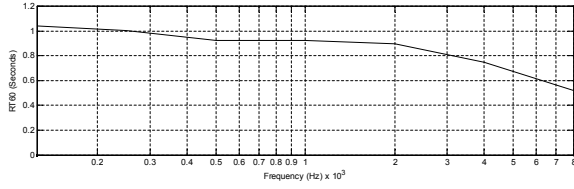


FIG. 2. Spatially averaged RT60 reverberation time over frequency for hall.

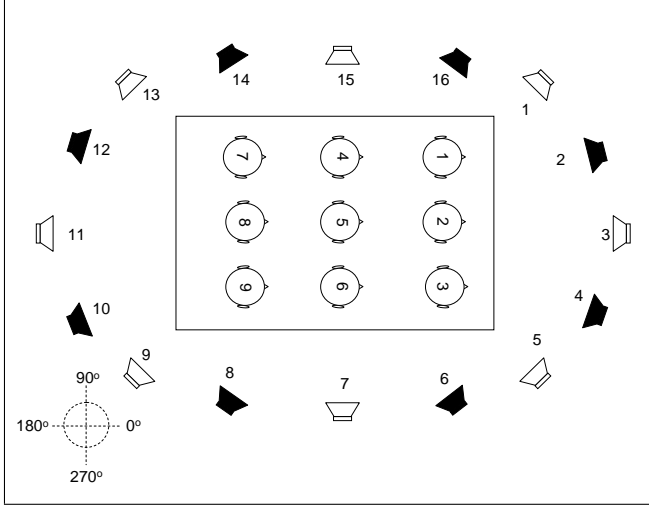


FIG. 3. Geometry of loudspeaker array and audience area for monophonic listening tests.

recordings of male speech, female speech, Gaussian white noise and music with fast transients.

Most spatialization techniques are optimised for circular loudspeaker arrays where each loudspeaker is equidistant from the centre listening position. Therefore, when utilizing irregular arrays, appropriate gain and delay adjustments must be implemented. This is generally required for lateral loudspeakers (speakers 7 and 15 in this case) which are positioned closer to the centre position; an inevitable consequence of attempting to place a circular array in a rectangular room. The appropriate delay was applied to each of the two lateral loudspeakers when encoding the test signals. The gain adjustments were applied to these two loudspeakers by calibrating each loudspeaker in the array to 70 dBA at the centre listening position. This approach is preferable to using the inverse square law when operating in a reverberant acoustic environment, due to the superposition of the direct and reverberant sound affecting the total SPL. Each loudspeaker axis line was positioned coincident with the centre listener position, and the speaker heights were set at 1.2m to tweeter.

B. Subjective Test Procedure

A series of listening tests was undertaken using groups of nine test subjects. Each group was presented with vir-

tual and monophonic sources from pseudorandom (predetermined) positions located about the speaker array and were then asked to identify the location of the sources via a questionnaire running concurrently with the tests. This randomized method was used to negate any order effects during the tests. In order to assess the effect of various stimuli, users were presented with 1 second unfiltered recordings of male speech, female speech, Gaussian white noise and music with fast transients. Each sample was presented twice, followed by a short interval before the next presentation. Listeners were asked to keep their heads in the forward direction throughout the test. Upon completion of one iteration of the test each listener was asked to move to the next seat for another randomised iteration. Each of the listeners' answers were weighted, depending on the confidence level of the listener with their choice, with weightings of $1/n$, where n is the number (or range) of speakers that a listener felt the sound originated from. From this, a histogram $\{h(\theta_i)\}_{i \in [1:16]}$ collecting all the listeners' answers for stimuli from each source location was computed for each seat. The angular mean $\bar{\theta}$ and the unbiased standard deviation σ_θ at each listener position are computed:

$$\bar{\theta} = \frac{\sum_{i=1}^{16} h(\theta_i) \cdot \theta_i}{\sum_{i=1}^{16} h(\theta_i)} \quad (1)$$

$$\sigma_\theta = \sqrt{\frac{\sum_{i=1}^{16} h(\theta_i)(\theta_i - \bar{\theta})^2}{(\sum_{i=1}^{16} h(\theta_i)) - 1}} \quad (2)$$

C. Objective Assessment Procedure

Binaural recordings can provide an objective measure of localization accuracy to support the findings of the perceptual listening tests. While both ITD and ILD cues can be calculated from binaural recordings, accurate estimates of azimuthal angle are difficult to infer from computation of the ILD in a reverberant environment. This is due to the fact that the superposition of the room reflections at the ear gives rise to significant differences of ILD measurements to that of measurements taken in a free-field environment for the same data window lengths. The ITD, however, is a more reliable estimate in this regard and is calculated using the normalized interaural cross correlation function (IACF) with the left and right ear signals, $x_1(t)$ and $x_2(t)$, given by,

$$IACF(\tau) = \frac{\int_{t_1}^{t_2} x_1(t)x_2(t+\tau)dt}{\sqrt{\int_{t_1}^{t_2} x_1^2(t)dt \int_{t_1}^{t_2} x_2^2(t)dt}} \quad (3)$$

The $IACF$ has a range of $[-1,1]$ which gives a measure of the correlation between the received signals in the integration limits t_1 to t_2 as a function of the time delay τ . Therefore the absolute value of the $IACF$ is a maximum when τ equals the true delay between $x_1(t)$ and $x_2(t)$, i. e. ,

$$T = \arg(\max_{\tau} |IACF(\tau)|) \quad (4)$$

A sampling rate of 96kHz was utilized throughout the measurements in order to maintain high resolution in the time delay estimation.

III. SOURCE LOCALIZATION ANALYSIS

Before the localization of virtual sources can be assessed, we need to first look at the localization of a real monophonic source under similar conditions. The localization performance for such *real* sources will provide a base measure of the best possible performance we can expect for a virtual source created using a multichannel spatialization scheme.

A. Localization of Monophonic Sources

The test results showed that excellent localization was achieved for frontal monophonic sources positioned at loudspeaker 2 for all nine seat positions. Both music and speech sources were localized to the correct loudspeaker with little deviation. The results for a white noise source were similarly good, albeit with some small deviations for the results at seat 6. The results for white noise and music sources originating at loudspeaker 6 can be seen in Figure 4. The mean results are all within 5° of the presented angle indicating that reasonably good localization was achieved at most seat positions. The level of deviation exhibited at seats 1, 2, 7, 8 and 9 is not insignificant and suggests that the room acoustics have some impact on localization accuracy. Comparable results were achieved for male and female speech sources positioned at this loudspeaker. The results for male and female speech sources positioned at loudspeaker 10 are shown in Figure 5. The best localization was achieved for male speech with all of mean angles falling within 10° of the presented angle. The results for female speech, white noise and music sources were broadly comparable and exhibit some deviation and mean errors. In general, the results are slightly worse than for a frontally biased lateral source but still reasonably accurate. The results for a rear-lateral source positioned at loudspeaker 14 are similar to those found for a rear source. However, while these results display greater deviation and fewer matching mean angles than for frontal source positions, reasonable localization was achieved with most subjects localizing the source to within 10° of the presenting loudspeaker.

These results suggest that for most combinations of listening and source position, the localization blur is not sufficiently strong to cause a listener to localize a monophonic source to the wrong location when using a non-circular 8-loudspeaker array. However, for extreme cases such as a front-corner listening position with a source positioned to the rear, then correct localization cannot be guaranteed.

Binaural recordings were taken at each listener point in the hall for broadband white noise and the ITDs were computed for each source location using the IACF. Figure

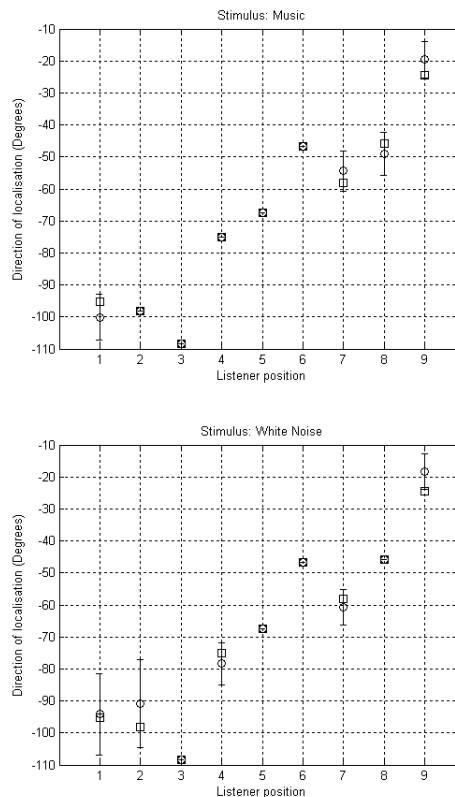


FIG. 4. Subjective localization of monophonic source stimuli presented at loudspeaker 6 for all listener positions. $\circ = \theta$, $\square = \sigma_T$, \pm = $\pm\sigma_\theta$

6 shows the time delay estimates for a white noise source located at loudspeakers 2, 6, 10 and 14.

The calculated ITDs represent the direction of the monophonic sources, where a negative delay indicates a source to the left of the head, and a positive delay represents a source at the right. It is important to remember that due to the cone of confusion associated with the ITDs, there are two possible azimuthal source angles that can be associated with the time delay: one in the frontal plane, and one in the rear plane¹. It is also noted that the accuracy of the delay estimation is dependent on the spectral content of the source. Figure 7 shows the IACF for a source at loudspeaker 2 for listener position 1. It can be seen that as the frequency content is increased in $1/3$ octave bands, the maximum peak in the correlation becomes more defined. It can be concluded then, as is the case with subjective localization⁶, that the more spectrally dense the signal, the better the estimate of localization. In our objective analysis it is therefore practical for us to assess the systems using broadband noise. Thus, the time delay estimates for white noise given in Figure 7 represent base ITD measurements for assessing the localization accuracy of the spatialization systems.

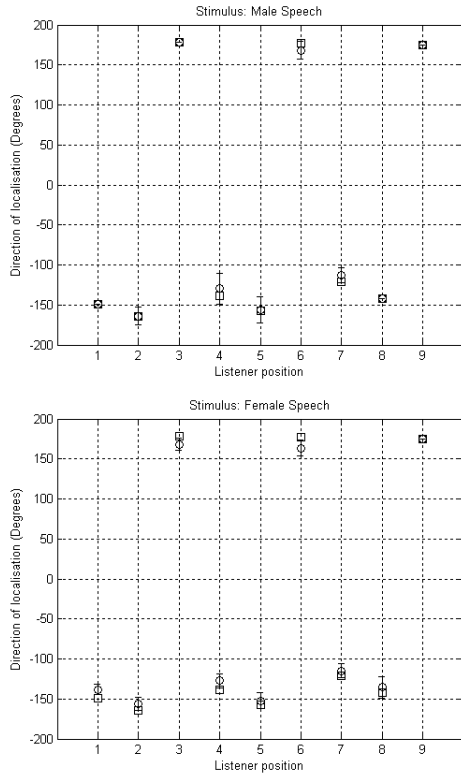


FIG. 5. Subjective localization of monophonic source stimuli presented at loudspeaker 10 for all listener positions. $\circ = \hat{\theta}$, $\square = \theta_T$, $\pm = \pm\sigma_\theta$

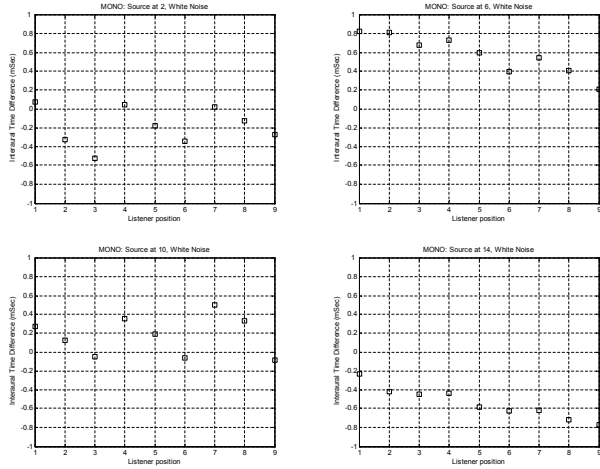


FIG. 6. ITD estimation for white noise presentations from loudspeakers 2, 6, 10 and 14.

B. Subjective Localization of Virtual Sources

Selected results for the virtual source localization tests are shown in Figures 8 and 9. No significant variations were found for different source stimuli, hence only the results for male speech are presented here.

With the Second Order Ambisonics system, the four loudspeakers surrounding the virtual source position are

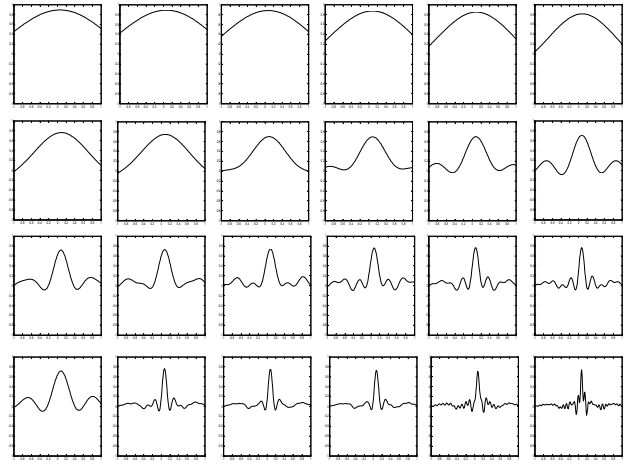


FIG. 7. Increase in IACF accuracy with cumulative 1/3 octave bands.

the dominant contributing loudspeakers in the array. The results for a frontal male speech presentation in Figure 8(a) clearly illustrate this trend as listeners at seat 1 consistently localized the source to loudspeaker 1, while listeners at seat 2 consistently localized the source to loudspeaker 3. Similarly, listeners at seats 3 and 4 consistently localized the source to the other two main contributing loudspeakers in the array, namely loudspeakers 5 and 15. The best localization was achieved at the rear centre listening position indicating that localization accuracy increases with distance from the source position, and hence the contributing loudspeakers. A similar bias is also evident for the other source positions with the worst localization being experienced at the listening positions closest to the source. In general, for lateral sources, the best localization was achieved at the listener positions at the opposite side to the source as can be seen in Figure 9(a). Interestingly, the results for a rear source positioned at loudspeaker 10 are at least comparable, and slightly better for the centre listening position, than the results for a frontal source. These results imply that localization accuracy is seriously degraded with Second Order Ambisonics systems for non-central listeners seated close to the array. Nearfield effects in Ambisonics were predicted by Gerzon long before the development of higher order systems and recent research by Daniel²⁰ has also confirmed the influence of these effects on localization accuracy. However, the significant deviations from $\hat{\theta}$ at the centre listening positions suggest that other factors are also influencing the perceived localization.

The localization results for Spat B-format Ambisonics encoding display a similar trend as for the higher order system with localization being consistently biased toward the nearest contributing loudspeaker. However this bias is even more exaggerated with B-format encoding and the localization accuracy decreases correspondingly. This can be clearly seen in the results for a frontal source in Figure 8(b). The results for seats 1, 2 and 3 are similar to the higher order system but the results for seats 6 and

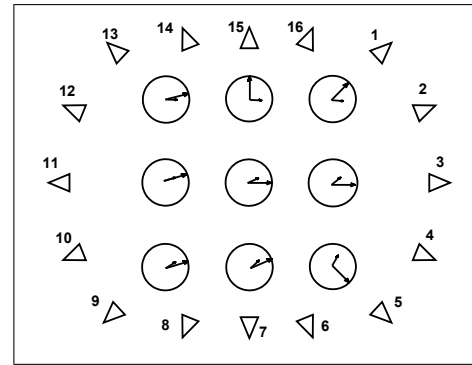
7 are much further away from the virtual source position. Similarly, for a rear-lateral source at loudspeaker 14, the results at seats 6 and 7 are worse than for the higher order system. The results for a rear and rear-lateral sources display a similar trend as can be seen in Figure 9(b). The results at the central listening position are slightly improved for the B-format system, particularly for front and front-lateral sources. These results seem to confirm that B-format Ambisonics functions better for a single, centrally positioned listener than higher order systems. However the localization accuracy for non-central listeners is improved somewhat when using higher order systems.

The results for a presentation of male speech positioned at loudspeakers 2 and 14 using VBAP are presented in 8(c) and 9(c). Again, no significant variations were found for different source stimuli. The results demonstrate similar nearfield biases to those reported for both Ambisonic systems. However, due to the number of contributing loudspeakers with VBAP (a maximum of two), smaller deviations from θ_T were found. The results for a frontal source clearly illustrate this nearfield bias as at seat positions 1 and 4 the source is consistently localized to loudspeaker 1, while at seat positions 2, 3, 5 and 6 the source is consistently localized to loudspeaker 3. The best localization was achieved at the seats furthest away from the source and at the same side of the hall, namely seats 7 and 8. The results for a rear source display similar trends except for the middle row, where the localization improved somewhat. Similar results were obtained for both lateral sources with the best localization being achieved at seat positions furthest away from the source. Also with lateral sources, localization increasingly collapses to the nearest contributing loudspeaker with decreasing distance to the virtual source.

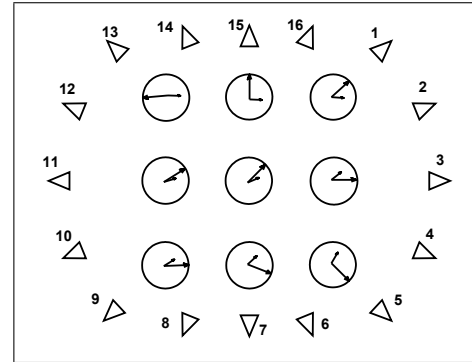
The localization results for a presentation of male speech positioned at loudspeakers 2 and 14 using DSS are shown in 8(d) and 9(d). Once again, no significant variations were found for different source stimuli. The results are very similar to those of VBAP, with localization being mainly biased towards the nearest contributing loudspeaker. However, the results for DSS show significantly greater variance about the mean than VBAP, especially at the seating positions furthest away from the source. This is not surprising considering that there are a greater number of contributing loudspeakers with DSS than with VBAP. It should be noted, however, that each $\bar{\theta}$ compares favourably with VBAP for these listener positions. The results for lateral sources correlate well with those for VBAP with greater deviations about $\bar{\theta}$.

C. Objective Localization of Virtual Sources

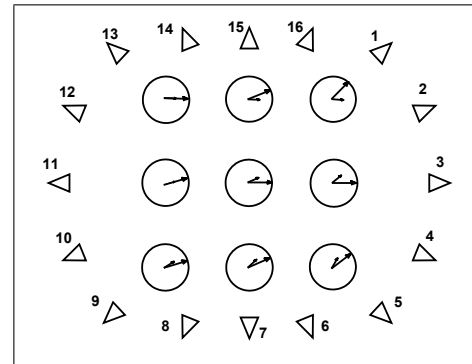
Binaural recordings of white noise were taken at each listener point in the hall for the same angular presentations as the subjective experiments. From these recordings the ITDs were inferred for each presentation. A comparison of all systems for localization of white noise from loudspeakers 2 and 14 is shown in Figures 10 and 11 respectively.



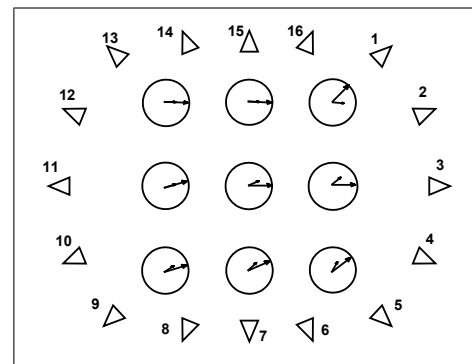
(a) Ambisonics



(b) Spat

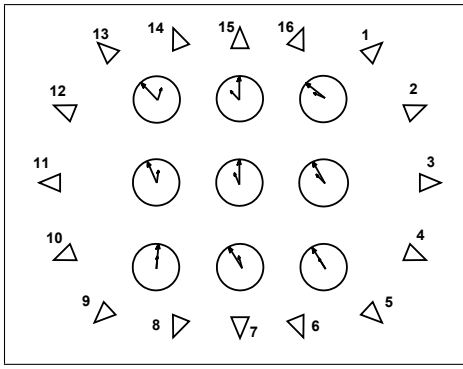


(c) VBAP

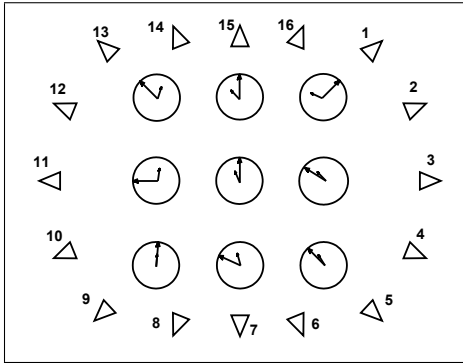


(d) Delta Stereophony

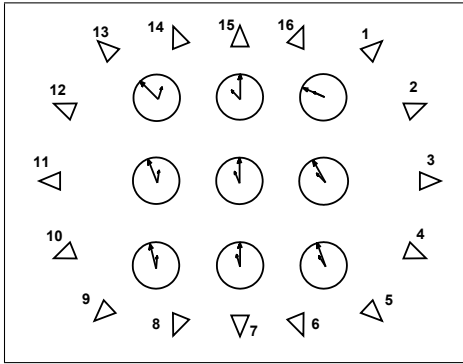
FIG. 8. Localization angles for a source at speaker 2 with male speech. The longer arrow represents the angular mean $\bar{\theta}$ computed at each listener position while the shorter arrow indicates the true angle θ_T to the presented source position.



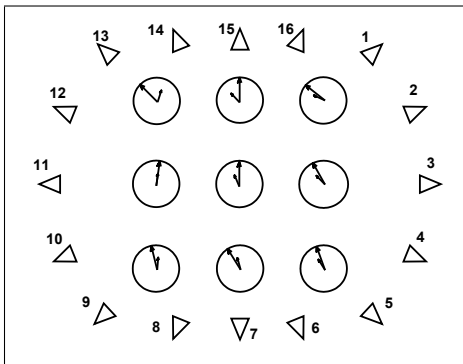
(a) Ambisonics



(b) Spat



(c) VBAP



(d) Delta Stereophony

FIG. 9. Localization angles for a source at speaker 14 with male speech. The longer arrow represents the angular mean $\bar{\theta}$ computed at each listener position while the shorter arrow indicates the true angle θ_T to the presented source position.

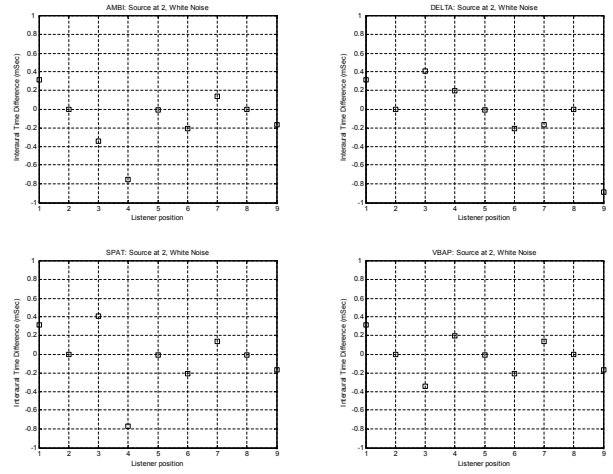


FIG. 10. ITD estimation for a white noise virtual source located at loudspeaker 2.

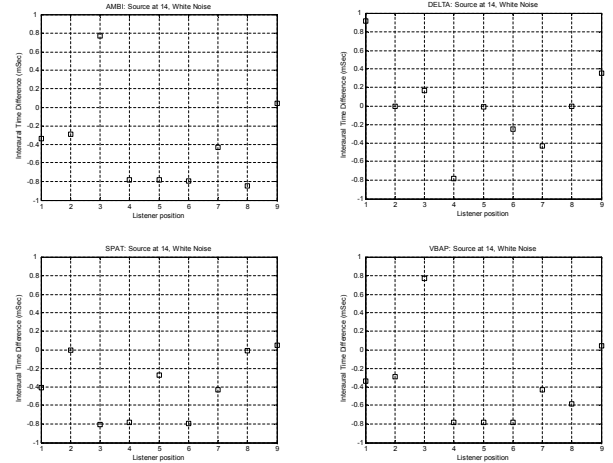


FIG. 11. ITD estimation for a white noise virtual source located at loudspeaker 14.

For Second Order Ambisonics biases are exhibited in the direction of the closest contributing loudspeakers to the virtual source for each position in the array. In support of the perceptual results we note that the ITDs match best with the monophonic sources at seats 5 and 6. It is interesting to note that there are also biases exhibited at seat 5, even after system calibration. For sources presented at loudspeaker 2, there are consistent ITDs indicating that the source is in the direction of loudspeaker 3. For sources at loudspeaker 6, there is a consistent bias towards loudspeaker 5, and for sources at loudspeaker 10, there is consistent bias towards loudspeaker 11. These results support the findings of the perceptual tests that consistent accurate localization is not achieved, even at the centre position.

The ITDs for Spat repeat the same trends of localization away from the virtual source. Lateral source cues are comparable to second-order Ambisonics, and the rear localization cues are mainly compromised. VBAP exhibits

similar ITD biases to Ambisonics for frontal localization, generally away from the source and towards the physical loudspeaker positions. The system performs adequately in terms of rear localization. Delta Stereophony exhibits slightly better frontal source localization for the rear listener positions, but compromised localization for rear sources. Its lateral localization performance is comparable to the other systems.

D. Observations on IACF objective analysis

An interesting observation in the calculation of the IACF for images created from multiple sources is that due to the complex interaction of loudspeaker signals at the ears, several significant peaks occur within the correlation window. Figure 12 illustrates this, by comparing the IACF estimate for white noise at listener position 5 for both virtual and real sources. The spatialization system used to generate the virtual source is VBAP. We can clearly see here the presence of multiple delayed versions of the signal within the correlation window. The peak at the dashed line (-0.19mSec) shows the correct time delay estimate for the monophonic source. However, for VBAP localization is dominated by another peak at 0.02mSec. The existence of other peaks in the correlation can be

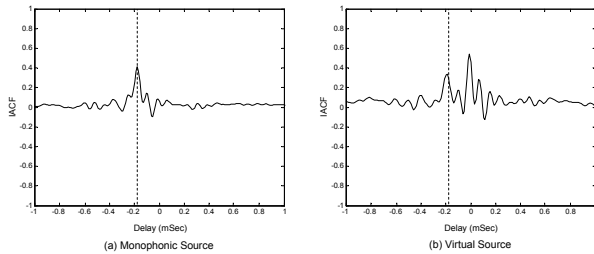


FIG. 12. IACF comparison for white noise located at loudspeaker 14.

modelled if we consider the case for two contributing loudspeakers with the same source stimulus. Consider the signal received at the left ear as $x_1(t)$, and the right, $x_2(t)$, expressed as

$$x_1(t) = \alpha_1 s(t) + \alpha_2 s(t + p_1) \quad (5a)$$

$$x_2(t) = \alpha_3 s(t + p_2) + \alpha_4 s(t + p_3) \quad (5b)$$

where $s(t)$ is the source signal, α is the attenuation factor, and p_1 , p_2 , and p_3 is the source delays relative to the left ear. We can express the unnormalized correlation as

$$R_{12}(\tau) = \int_{-\infty}^{\infty} G_{x_1 x_2}(\omega) e^{j\omega\tau} d\omega \quad (6)$$

where

$$G_{x_1 x_2} = X_1(\omega) X_2^*(\omega) \quad (7)$$

where $*$ denotes the complex conjugate. Taking the Fourier Transform of $x_1(t)$ and $x_2(t)$ we get

$$X_1(\omega) = \alpha_1 S(\omega) + \alpha_2 e^{-j\omega p_1} S(\omega) \quad (8a)$$

$$X_2(\omega) = \alpha_3 e^{j\omega p_2} S(\omega) + \alpha_4 e^{-j\omega p_3} S(\omega) \quad (8b)$$

Thus

$$\begin{aligned} G_{x_1 x_2}(\omega) &= \alpha_1 \alpha_2 e^{-j\omega p_2} S(\omega) S^*(\omega) \\ &\quad + \alpha_1 \alpha_4 e^{-j\omega p_3} S(\omega) S^*(\omega) \\ &\quad + \alpha_2 \alpha_3 e^{j\omega p_1} e^{-j\omega p_2} S(\omega) S^*(\omega) \\ &\quad + \alpha_2 \alpha_4 e^{j\omega p_1} e^{-j\omega p_3} S(\omega) S^*(\omega) S(\omega) S^*(\omega) \end{aligned} \quad (9)$$

This then gives in the time domain

$$\begin{aligned} \mathcal{F}^{-1}[G_{x_1 x_2}] &= \int_{-\infty}^{\infty} \alpha_1 \alpha_3 S(\omega) S^*(\omega)^* e^{j\omega(\tau - p_2)} d\omega \\ &\quad + \int_{-\infty}^{\infty} \alpha_1 \alpha_4 S(\omega) S^*(\omega) e^{j\omega(\tau - p_3)} d\omega \\ &\quad + \int_{-\infty}^{\infty} \alpha_2 \alpha_3 S(\omega) S^*(\omega) e^{j\omega(\tau - p_2 + p_1)} d\omega \\ &\quad + \int_{-\infty}^{\infty} \alpha_2 \alpha_4 S(\omega) S^*(\omega) e^{j\omega(\tau - p_3 + p_1)} d\omega \\ &= \alpha_1 \alpha_3 \delta(\tau - p_2) + \alpha_1 \alpha_4 \delta(\tau - p_3) \\ &\quad + \alpha_2 \alpha_3 \delta(\tau - p_2 + p_1) \\ &\quad + \alpha_2 \alpha_4 \delta(\tau - p_3 + p_1) \end{aligned} \quad (11)$$

Thus, we see how in this simple case of two loudspeakers, multiple correlation peaks arise due to the superposition of the loudspeaker signals at the ears. In the ideal anechoic case, where the loudspeakers are symmetrical about the listener, this should not pose a problem and the dominant correlation peak is at a time delay comparable to that of a real monophonic source. However, cases can arise where the source signals arrive at the ears out of phase, resulting in two equally dominant correlation peaks, neither of which reflects the intended source direction. Such conditions have been documented by Okano²¹, as well as how strong lateral sources can have a negative effect on localization accuracy. Thus, since the IACF model implemented here chooses only one maximum peak in the correlation, the presented results may not always reflect that of the perceived source angle. Further work is required into a model for the perceptual analysis of multiple source peaks, that also looks outside the cross head delay window, to accommodate early delays that could potentially suppress the precedence of the initial wavefronts. As a precursor to this work, the authors have conducted preliminary investigations into the use of pre-filters applied prior to the normalization of (3) in the frequency domain, so as to enhance the observed peaks for analysis of multiple time delays. We investigated the Phase Transform (PHAT) processor as presented by Knapp and Carter²² in this regard. The results of these tests, which will be presented at a later date, show that the processor improves the time delay estimation of differing source stimuli by effectively weighting the phase function of the received signals at the ears uniformly over the entire frequency band, leading to narrower peaks in the time delay estimation as shown in Figure 13.

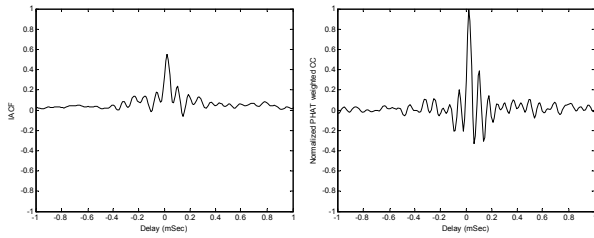


FIG. 13. Comparison of IACF to PHAT weighted cross correlation for monophonic white noise presentation from speaker 2 at seat 7.

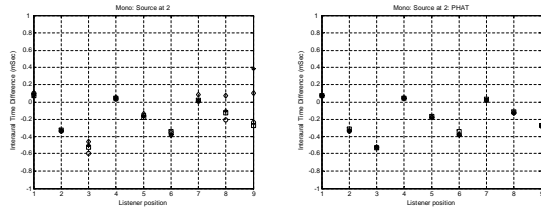


FIG. 14. Comparison of IACF to PHAT weighted cross correlation for various source stimuli for monophonic loudspeaker 2 presentation. \circ = Female Speech, \diamond = Male Speech, \square = White Noise, $*$ = Music.

IV. OVERALL PERFORMANCE

In the preceding sections we analysed the localization performance of each spatialization technique for each source position. However, it is also important to find a measure of which spatialization technique provides the best overall performance. For this purpose, each system was analyzed in terms of its subjective *hit rate*, calculated by correlating the ideal localization histogram with the observed results. This measure, displayed in Figure 15 expresses the percentage localization accuracy over all source positions and stimuli.

As we can see, the localization performance of each system does not achieve that of monophonic sources. Overall the intensity panning systems perform better for front and back sources, with VBAP providing a 12.7% higher localization accuracy over DSS for rear sources. Higher Order Ambisonics performs consistently better than B-format at all source positions and its performance is comparable to VBAP for lateral rear sources. Interestingly, we see that even though lateral rear monophonic localization falls by over 24%, the localization accuracy of the intensity based systems do not fall by the same degree, and in fact both Ambisonics systems exhibit a slightly higher degree of accuracy for lateral sources over frontal sources.

V. DISCUSSION

The results presented in the previous sections indicate that neither intensity panning or Ambisonics techniques can create consistently localized virtual sources for a distributed audience in a reverberant environment. Source localization for non-central listener positions is

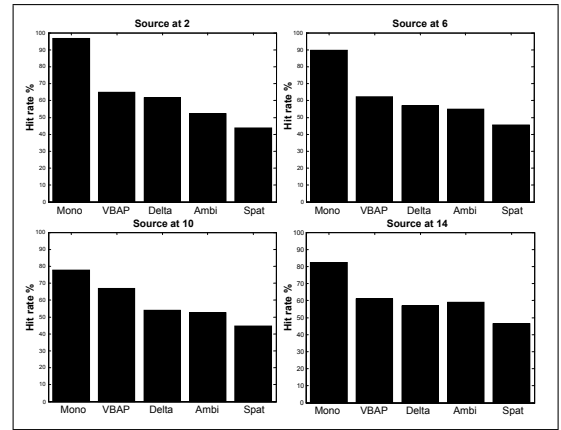


FIG. 15. Overall subjective localization performance.

consistently biased toward the nearest contributing loudspeaker in the array, irrespective of the spatialization technique or the nature of the source stimulus. Due to the number of loudspeakers used by the B-format and Second Order Ambisonics systems (predominantly four loudspeakers) this bias results in a significant range of perceived source angles at different listener positions. In both cases, the localization accuracy increases with distance from the source, which again illustrates this problem. The results at the centre listening position were generally more consistent for varying source positions than for non-central listening positions. However, even at this preferred listening position localization errors of between 10° and 20° regularly occurred. These results suggest that the room acoustics are also having a significant effect on localization accuracy and further research carried out by the authors supports these findings²³. B-format Ambisonics performed slightly better at the centre listening position but the higher order system produced better results at non-central listening positions. These findings support the view that B-format is preferable for a single listener while higher order systems are more suitable for a presentation to a distributed audience.

The results for both VBAP and DSS illustrate the well known limitations of the stereophonic principle for off-centre listening positions. As with Ambisonics, both of these systems displayed biases towards the nearest contributing loudspeaker. However due to the smaller number of loudspeakers with VBAP, this did not affect localization to the same extent. The results for DSS were largely comparable to those of VBAP albeit with a slightly increased deviation for distant listener positions due to the large number of contributing loudspeakers. However, this effect is countered somewhat by the more consistent SPL levels created across the entire listening area with this system.

VI. CONCLUSION

In this paper we have assessed the subjective and objective performance of various spatialization techniques in terms of their localization accuracy for a distributed

audience in a reverberant environment. The results of a series of listening tests were presented and these findings were supported by calculated ITDs inferred from high resolution binaural measurements recorded in the test environment. The results for monophonic source localization indicate that a distributed audience can generally well localize a real source in a reverberant environment. However, the comparison of the results for monophonic and virtual source localization suggest that if consistent and accurate source localization is required, then virtual sources created with these systems cannot be relied upon. The results for B-format ambisonics show that source localization with this system is severely compromised at non-central listening positions with localization being consistently biased toward the nearest contributing loudspeaker. The extension of Ambisonics to higher orders has been shown to improve localization for non-central listeners however these systems perform slightly worse at the centre position. Both DSS and VBAP systems also suffered from a similar localization bias toward the nearest contributing speaker. However, as VBAP only ever utilizes a maximum of two loudspeakers at any one time, localization accuracy was not degraded to the same extent with this system. The intensity panning variation of DSS implemented in these tests does provide a more uniform SPL coverage over the listening area, however due to the greater number of contributing loudspeakers used, the localization accuracy of this system is less than that of VBAP.

Systems that attempt to recreate the entire soundfield over an extended listening area could potentially overcome these problems. The Wavefield Synthesis (WFS) concept developed by Berkhout at TU Delft²⁴ is one such system and further research is required to assess its ability to recreate the accurate wavefronts required for consistent source localization across a large listening area. The ability of such systems as VBAP, DSS, Ambisonics and WFS to recreate dynamically moving sources also needs to be investigated as the capability of these systems in this regard is also of significant importance.

Acknowledgments

The authors wish to thank all the participants of the listening tests and in particular the postgraduate students of the Department of Electronic and Electrical Engineering and the Music and Media Technology course, Trinity College Dublin. The assistance of Mr. Damien Kelly is also greatly appreciated.

- ¹ J. Blauert, *Spatial Hearing* (MIT Press Cambridge MA) (2003).
- ² J. Rayleigh, *Theory of Sound* (Dover, N.Y.) (1945).
- ³ E. A. MacPherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The Duplex Theory of sound localization revisited", *Journal of the Acoustical Society of America* **111**, 2219–2236 (2002).
- ⁴ S. G. Weinrich, "Horizontal Plane Localization Ability and Response Time as a Function of Signal Bandwidth", in *Au-*

- dio Engineering Society Preprint 4007; AES Convention 98* (1995).
- ⁵ T. T. Sandel, D. C. Teas, W. E. Feddersen, and L. A. Jeffress, "Localization of Sound from Single and Paired Sources", *Journal of the Acoustical Society of America* **27**, 842–852 (1955).
- ⁶ W. M. Hartmann, "Localization of sound in rooms", *Journal of the Acoustical Society of America* **74**, 1380–1391 (1983).
- ⁷ B. Rakerd and W. M. Hartmann, "Localization of sound in rooms, II: The effects of a single reflecting surface", *Journal of the Acoustical Society of America* **78**, 524–533 (1985).
- ⁸ M. A. Gerzon, "Periphony: With-height sound reproduction", *Journal of the Audio Engineering Society* **21**, 2–10 (1973).
- ⁹ M. A. Gerzon, "Ambisonics in multi-channel broadcasting and video", *Journal of the Audio Engineering Society* **33**, 859–871, (1985).
- ¹⁰ D. G. Malham and A. Myatt, "3-D sound spatialisation using Ambisonic techniques", *Computer Music Journal* **19**, 58–70 (1995).
- ¹¹ E. Benjamin, R. Lee, and A. J. Heller, "Localization in horizontal-only Ambisonic systems", in *121st Convention of the Audio Engineering Society* (2006).
- ¹² J. Schacher and P. Kocher, "Ambisonics spatialization tools for max/msp", www.icst.net (2006).
- ¹³ D. G. Malham, "Experience with large area 3-D Ambisonic sound systems", *Institute of Acoustics* **8**, 209–216 (1992).
- ¹⁴ V. Pulkki, "Virtual sound source positioning using Vector Base Amplitude Panning", *Journal of the Audio Engineering Society* **45**, 456–466 (1997).
- ¹⁵ J. Jot, "Real-time spatial processing of sounds for music, multimedia and human-computer interfaces", (1997), URL citeseer.ist.psu.edu/jot97realtime.html.
- ¹⁶ J. Chowning, "The simulation of moving sound sources", *Journal of the Audio Engineering Society* **19**, 2–6 (1971).
- ¹⁷ F. R. Moore, "A general model for spatial processing of sounds", *Computer Music Journal* **7**, 6–15 (1983).
- ¹⁸ W. Ahnert, "Complex Simulation of Soundfields by the Delta Stereophony System (DSS)", *Journal of the Audio Engineering Society* **35**, 643–652 (1987).
- ¹⁹ W. Ahnert, "Problems of Near-Field Sound Reinforcement and of Mobile Sources in the Operation of the Delta Stereophony System (DSS) and Computer Processing of the Same", in *82nd Convention of the Audio Engineering Society* (1987).
- ²⁰ J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format", in *23rd International Conference of the Audio Engineering Society* (2003).
- ²¹ T. Okano, "Image shift caused by strong lateral reflections, and its relation to inter-aural cross correlation", *The Journal of the Acoustical Society of America* **108**, 2219–2230 (2000).
- ²² C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", *IEEE Transactions on Acoustics, Speech and Signal Processing* **24**, 320–327 (1976).
- ²³ G. Kearney, E. Bates, D. Furlong, and F. Boland, "A Comparative Study of the Performance of Spatialization Techniques for a Distributed Audience in a Concert Hall Environment", in *31st International Conference of the Audio Engineering Society* (2007).
- ²⁴ A. J. Berkhout, "A holographic approach to acoustic control", *Journal of the Audio Engineering Society* **36**, 977–995 (1988).