# PSYCHOACOUSTIC EVALUATION OF SPATIAL AUDIO REPRODUCTION SYSTEMS

D. Satongar University of Salford, UK
C. Dunn BBC Research and Development, Salford UK
Y. W. Lam University of Salford, UK
F. F. Li University of Salford, UK

***Abstract*** - A key role of spatial audio reproduction systems is to recreate a sound field. In doing this, adequate psychoacoustic cues must be created for a listener. To assess this ability, the interaural time and level difference cues synthesised by Ambisonic and amplitude-panned reproduction systems are studied. An analytical sphere binaural model and a binaural model using measured HRTF data from the MIT KEMAR database were used to obtain ITD and ILD cues for the two reproduction methods with three loudspeaker layouts. By comparing cues for central and off-centre listening positions, the relationship between listening area 'sweet spot' and the low-level psychoacoustic cues has been identified. For central listening positions, results for the amplitude panned ITU 5.0 layout highlight large ITD errors with first-order Ambisonics showing improved results. Third-order Ambisonics was required to achieve improved ILD error over the ITU 5.0 system. For off-centre listening analysis, fourth-order Ambisonics on an octagon layout was required to achieve improved ITD and ILD error values over the ITU 5.0 system. The analytical limitations of using low-level psychoacoustic cues are discussed, with suggestions for further work in the area.

## 1 INTRODUCTION

The work in this paper presents the start of an ongoing investigation into developing new performance metrics for immersive reproduction systems. Currently the focus is towards the variability of listening area reproduction and how this correlates with horizontal Ambisonic reproduction variables, with broadcast applications in mind.

There are a number of objective localisation measures used in spatial audio reproduction assessment. These include the popular velocity and energy vectors, integrated difference error and instantaneous pressure values. However, although these give a good indication of the reproduced sound field accuracy, the lack of psychoacoustic representation can mean that certain perceptual attributes remain unexplained. For example, failing to model sound scattering from the human head is a significant omission[1]. Therefore this study focuses on the psychoacoustic cues created by a reproduced sound field.

Spatial audio reproduction systems aim to reproduce a perceived sound image and therefore induce realistic psychoacoustic localisation cues. Interaural time difference (ITD) and interaural level difference (ILD) measures are implemented due to both their well-formulated use in approximating perceived lateralisation[2] and also the ease of their implementation. Using a spatial audio model it was possible to approximate cues for both real point sources and phantom sources (synthesised by a number of speakers). The use of an analytical spherical head model also meant that this data could be approximated for a number of off-centre listening positions, meaning listening area variability could be approximated.

The paper can be split into two sequential stages:

1. The first part of the study focuses on the ability of the chosen reproduction systems to create phantom (panned) sources compared with real point sources for a ***central listening position***.

2. The same method is then used in order to calculate ITD and ILD error values for ***off-centre listening positions*** in an attempt to approximate listening area characteristics. This involves taking the data from stage 1 and approximating an "averaged error" and repeating for each listening position for both ITD and ILD.

# 2 SPATIAL AUDIO REPRODUCTION

Two spatial audio reproduction technologies were compared - Ambisonics, and vector base amplitude panning (VBAP)[3] on an ITU 5.0 layout as a reference. The reason different loudspeaker layouts were chosen for the two reproduction methods was to ensure that the commonly used reproduction combination (ITU 5.0 layout with amplitude panning) was used as a reference. Rendering Ambisonics over the ITU 5.0 system would have needed irregularity compensation and therefore introduced an addition variable; hence the hexagon and octagon layouts were used.

## 2.1 Ambisonics

Ambisonics is a periphonic spatial audio technology first developed in the 1970s[4].It has been popular among academic and research institutions since it's inception but has only seen limited commercial adoption[5]. Ambisonics has the benefit of scalability, where a higher-order representation also includes all lower-order harmonics, allowing fast and efficient down scaling of Ambisonic order; this feature is desirable for broadcast applications.

In this paper horizontal-only (pantophonic) reproduction is considered, where Ambisonic height components can be omitted. Two Ambisonic loudspeaker layouts were chosen for analysis, a hexagon without centre front [30° 90° 150° 210° 270° 330°] and an octagon with centre front [0° 45° 90° 135° 180° 225° 270° 315°].

***Encoding -*** Ambisonic encoding is a method of *synthesising* an Ambisonic sound field (as opposed to recording with an Ambisonic microphone). This is done by scaling a mono sound signal into B-Format channels using spherical harmonic weighting coefficients. The coefficients are a function of the source azimuth (and elevation for 3d). The higher the Ambisonic order, the greater the number of encoded channels.

***Decoding*** - The sound field is reconstructed from spherical harmonics by the decoder. This involves determining loudspeaker decoding coefficients appropriate for the encoded Ambisonic signals and loudspeaker layout. The method chosen to calculate the gains in this experiment was through the use of a basic (as opposed to in-phase or max $r_E$) "Pseudo-inverse" method. Although it has been shown[6] that maximum energy (max $r_E$) decoding can improve localisation under certain circumstances and in-phase decoding can improve off-centre listening variability[7,8], the basic decoder was chosen to limit the number of simulation variables, especially regarding the second stage of the experiment. The effect of other Ambisonic parameters including decoder design may be investigated in future work.

## 2.2 Vector Base Amplitude Panning with ITU 5.0 Layout

2D Vector Base Amplitude Panning is a method of amplitude panning a phantom audio source between loudspeakers. It is a geometric based panning law in which the loudspeaker gains are determined by the position of a desired phantom source between two or more loudspeakers. VBAP can be used on periphonic loudspeaker layouts, however in this situation (ITU 5.0) the panning law is similar to stereophonic panning laws in that only one or two loudspeakers are ever in use for any source azimuth position. Therefore VBAP is representative of the amplitude-panned systems currently used in broadcast.

Pulkki highlights the ability of VBAP to use the minimum number of loudspeakers necessary to reproduce a phantom source, contrasting with the tendency of Ambisonic methods for all loudspeakers to contribute to reproducing a phantom source. VBAP is also highlighted as being beneficial for irregular loudspeaker layouts, a common problem for Ambisonic systems and an active area of research.

# 3 MODELLING

In order to approximate the sound field perceived by a listener, a number of computational models had to be implemented. These models can be split into three sections:

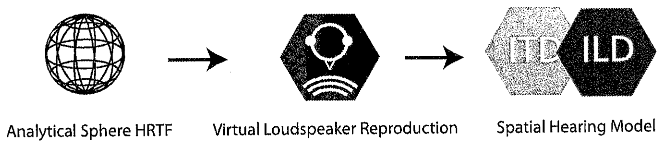

**Figure 1 - A block representation of the modelling methodology**

The experiment utilises two sets of Head Related Impulse Responses (HRIRs). The first uses data acquired from an analytical spherical head model and the second uses data from the MIT KEMAR database[9].

HRIRs corresponding to a specific spatial audio reproduction system are synthesised using a method discussed by Landone and Sandler[10] in which HRIRs from each loudspeaker (point source for the analytical model) are scaled relative to the gain of that loudspeaker for a particular phantom source azimuth. Summing these head related impulse responses at each ear, for all loudspeakers in the system, forms a global HRIR. ITD and ILD measures are then approximated from the global ear HRIRs.

## 3.1 Analytical Sphere

Blauert[2] has shown that the pressure on the surface of the sphere at locations similar to that of human ears gives a good approximation of pressure at the ears of a human head. The equation from Asvestas et al.[11] was used to approximate the velocity potential at any rotational point on the surface of the sphere due to a point source at distance $r_0$ (1.4m in this case) and angle $\theta = 0$.

Modelling the ears as being at 90° and 270° on the surface of the sphere, and calculating impulse responses for 0 - 355° with 5° intervals, the corresponding azimuth values were calculated to yield a dataset of impulse responses for the sphere model. The sphere radius was approximated[12] to 9cm.

A key benefit of the analytical model is the ability to make small changes to the source distances from the listener as the head is offset from the central listening position. This fine distance resolution is not available from most HRTF databases and would take a large amount of time to measure. The KEMAR HRIRs were used to validate the spherical head model results for a central listening position.

## 3.2 Low-Level Localisation cues

It was identified that the logical starting point for assessing perceptual horizontal localisation cues is through the use of the commonly used interaural time and level differences. The theory states that a sound source is lateralised by the human auditory system's ability to measure small time and level differences between the ears[2]. Both ITD and ILD data was approximated in 1/3-octave frequency bands.

### 3.2.1 Interaural Time Difference

There are many known methods of estimating ITD values and preferences vary between research disciplines. Neuroscience usually opts for a physiological approach such as place theory[13], whereas audio engineering adopts methods based on the comparison of HRIRs such as the maximum interaural cross-correlation coefficient (IACC) method firstly proposed by Kistler and Wightman[14]. The maximum IACC method was used due to its computationally simple implementation and also it's relatively good approximation of ITD[15]. This is performed by finding the argument (time delay) at which the IACC function of the two HRIRs reaches a maximum – the implementation of this is presented by Bates et al[16].

In order to validate the analytical sphere HRIR data, ITD data for real sources were calculated using both the KEMAR MIT database and the analytical sphere database. These can be seen below.

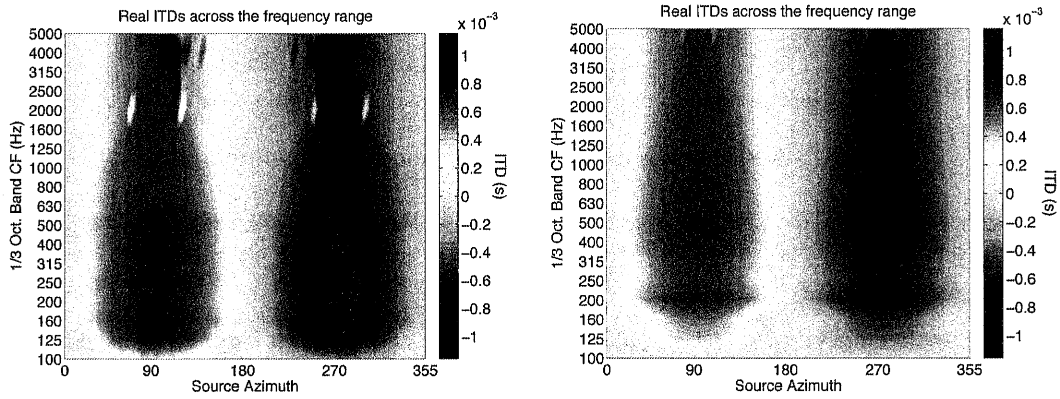

**Figure 2 – Real source ITDs using KEMAR MIT (left) and analytical sphere (right) data**

The plots highlight a strong similarity across frequency, azimuth and amplitude. One noticeable feature on the KEMAR data shows ITD values containing some anomalous peaks and troughs around azimuths of 70°, 110°, 250° and 290° in the frequency band around 2 kHz. This may be because the ITD approximation method relies on left and right ear HRIRs being well correlated, around the interaural axis the correlation is reduced[17]. Due to the simplification of the head and facial features to a simple sphere, this is not apparent in the sphere ITD data.

## 3.2.2 Interaural Level Difference

For ILD, the equation presented by Gaik[18] was used to calculate the energy difference between the two ears in dB from impulse responses. As with the ITD measurements, the sphere ILD simulations were compared to KEMAR measurements using a real source to validate the use of the analytical sphere model.

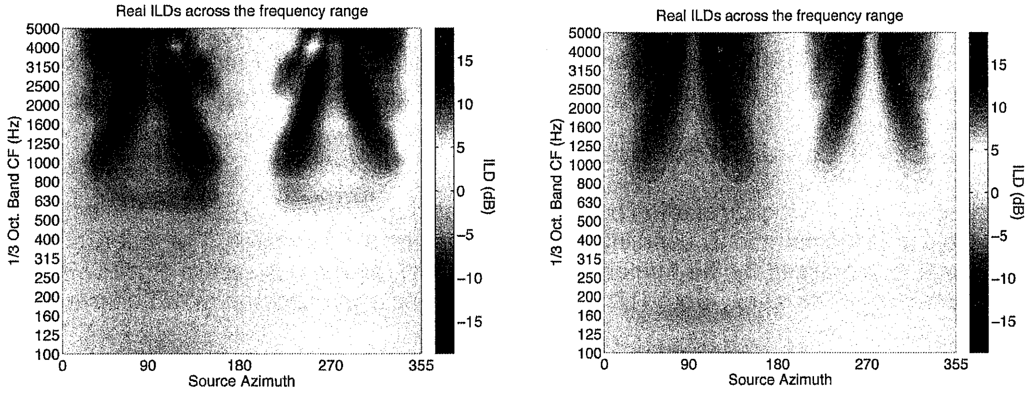

**Figure 3 – Real source ILDs using KEMAR MIT (left) and analytical sphere (right) data**

Some key points can be made regarding these results:

1. Plots are similar across all domains (azimuth, frequency band and ILD)
2. The KEMAR database contains 2-way symmetry between the left and right sides of the head whereas due it not containing facial features, the analytical sphere has 4-way symmetry.
3. Both data sets show the rear pressure dip at frequency bands above 1250 Hz, seen as a reduction in ILD magnitude at 90° and 270°. The dip is caused by pressure interference, with the KEMAR ILD results showing less dip in comparison to the spherical head model. Blauert[2] states that the reduced dip in the KEMAR data is due to the head resting on the neck, an attribute the analytical sphere model does not contain.

The analytical model achieves a reasonable approximation of the measured KEMAR ITD and ILD data and will be the data source for all simulation results presented below.

# 4 CENTRAL LISTENING POSITION

Our initial experiment considers localisation cues at the central listening position for a specified Ambisonic layout and compares these with cues from the ITU 5.0 layout rendered using VBAP.

ITD and ILD data were calculated for real and phantom sources synthesised by the different reproduction systems. The magnitude difference between these values represents an error metric in seconds for ITD and dB for ILD.

Using the 3-stage hearing model described above, ITD and ILD error values were computed by taking the magnitude difference in ITD/ILD values between a ***real source*** and a ***phantom source*** for each angle between 0° and 355° with 5° resolution. This was performed for each 1/3-octave frequency band. Colour intensity plots were created to show how the ITD and ILD error profiles change for Ambisonic order in comparison to the ITU 5.0 reference system. It is worthwhile to note that the ITD value is now an error magnitude so only contains positive values. Black squares highlight speaker locations.

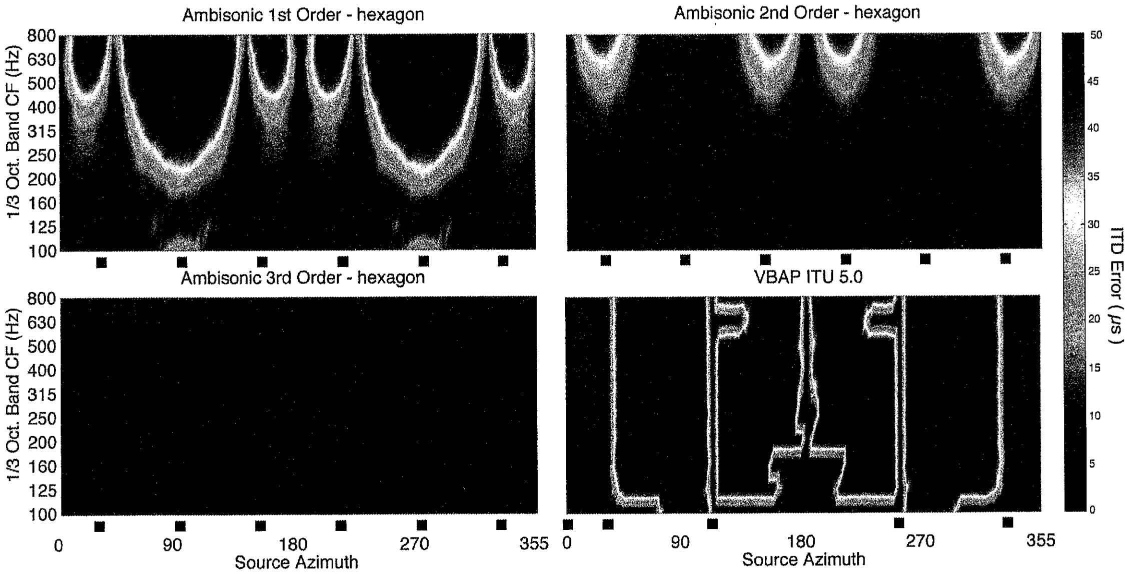

**Figure 4 – ITD error maps for Hexagon layout Ambisonic system and VBAP ITU 5.0 system.**

The ITD error maps shown in Fig. 4 highlight some key attributes:

1. Ambisonic ITD error magnitude is reduced as order increases
2. Spatial error distributions are related to speaker locations for the ITU 5.0 system

3. ITD error values for 3rd order Ambisonic are small (<15μs)

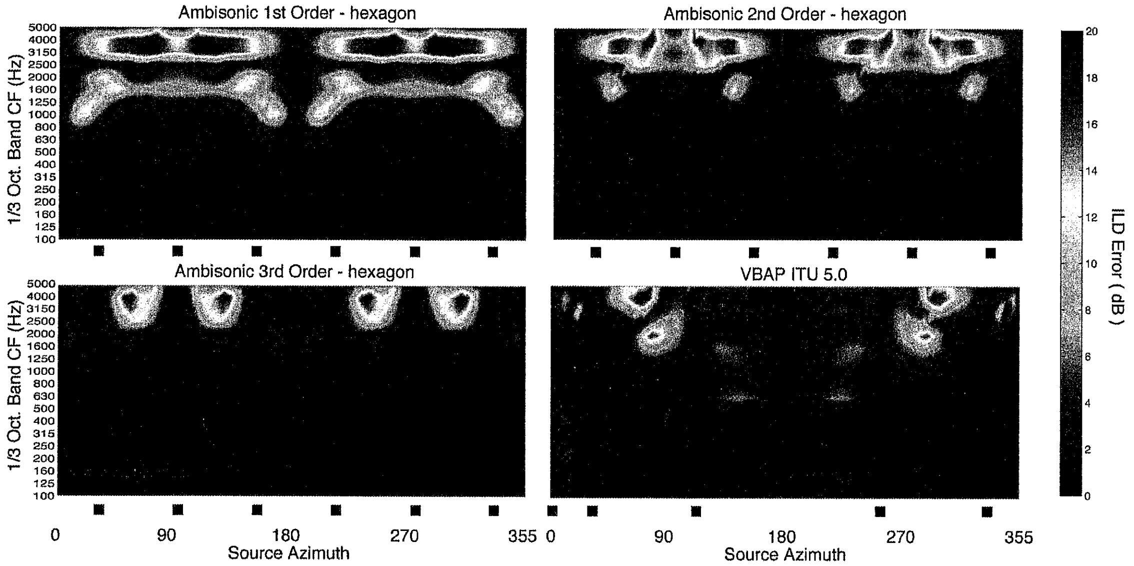

**Figure 5 – ILD error maps for Hexagon layout Ambisonic system and VBAP ITU 5.0 system.**

Some key attributes of the ILD error plots:

1. ILD errors reduce as Ambisonic order increases
2. Reduced ILD error is observed in the lower frequency bands
3. ITU VBAP 5.0 spatial error distribution remains related to loudspeaker locations
4. All results show a frequency dependency across source azimuth, behaviour also observed by Pulkki and Hirvonen[19]

Pulkki and Hirvonen also presented results for ITDA and ILDA (angular extrapolation of ITD and ILD) for first and second order Ambisonic octagon layout and the ITU 5.0 layout with pair-wise panning. Although different methods were used, some similarities can be seen with our simulation results.

- For first- and second-order Ambisonics, Pulkki and Hirvonen's results show large frequency dependency highlighted by an increase in low-frequency accuracy for second order.
- For the ITU layout, Pulkki and Hirvonen highlight increased consistency across the frequency range, which is also observed in our results.

Another assessment that can be made on the data presented above is to take a mean average across the whole error plot for each system, i.e. averaged across frequency and source azimuths. This will give some indication of the Ambisonic order needed to match VBAP ITU 5.0 performance for a central listening position.

| System | Mean ITD error (μs) | Mean ILD error (dB) |
|---|---|---|
| **1st Order** | 32 | 3.2 |
| **2nd Order** | 6.6 | 2.2 |
| **3rd Order** | 1.4 | 1.4 |
| **VBAP ITU 5.0** | 92 | 1.8 |

**Table 1 – Mean ITD and ILD error results for systems under test.**

The results highlight some key points regarding the specific comparison between a 6-speaker Ambisonic array and the VBAP ITU 5.0 system, for a listener at the central listening position:

- 3rd order Ambisonics is required to match the VBAP ITU 5.0 performance for ILD reproduction.
- 1st order Ambisonics gives improved ITD performance over VBAP ITU 5.0.

For the ILD simulations it is clear to see that as the Ambisonic order increases, the frequency band at which error becomes significant also increases. This is a feature identified by Daniel et al.[20] in which they define the upper frequency limits of accurate reconstruction. The table below shows the results of their 20% D-error frequency threshold against the frequency band at which ILD error reaches above 5 dB in our simulations. Noting that the simulation results presented in this experiment are centred on 1/3-octave bands, these show a clear similarity in trend with results recorded by Daniel et al.

| System | Daniel et al. 20% D-Error (Hz) | 5 dB error freq. band (Hz) |
|---|---|---|
| **1st Order** | 742.9 | 800 |
| **2nd Order** | 1345.8 | 1600 |
| **3rd Order** | 1959.5 | 2500 |

**Table 2 – Comparison of Ambisonic frequency error thresholds.**

# 5 OFF-CENTRE LISTENING POSITIONS

For the second stage of the experiment, the same method used to calculate virtual HRIRs for the central listening position was repeated at multiple off-centre listening positions. Using the analytical sphere model meant that the off-centre HRTFs could be quickly simulated for a number of different listening positions, without the need to make time-intensive dummy head measurements for the off-centre positions.

In order to approximate the performance of a listening area, data for each listening position (ITD and ILD error for frequency bands across the source azimuth range) were processed to achieve average ITD and ILD error values; this included B-Weighted (to approximate the frequency weighting of human hearing[21]) magnitude averaging across the frequency range, and equally-weighted magnitude averaging across source azimuths. ILD data was averaged across the full 1/3 octave frequency band range but ITD was only considered for frequency band values between 100 - 800 Hz.

The listening area analysis was only conducted for Ambisonic systems, using an octagonal layout for this part of the experiment rather than the hexagonal array used earlier.

It is important to try and consider how these errors relate to human perception. Therefore, values must be defined which indicate a realistic threshold of human perception on the ITD and ILD error scales. It is shown by Moore and Wakefield[22] that at angles between 60° and 100° the perceptual azimuth resolution is around 7°. Extrapolating from the KEMAR plots, this corresponds to around 90 μs ITD and around 2.5 dB ILD; these thresholds will be used for estimating listening area quality. The areas at which the errors are below these thresholds are highlighted on the plots shown in Fig. 6 and 7 with a red line.

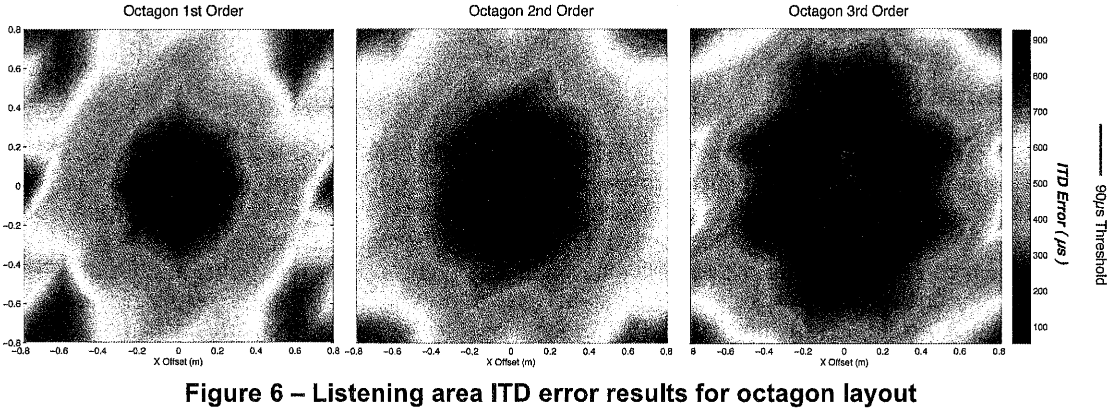

**Figure 6 – Listening area ITD error results for octagon layout**

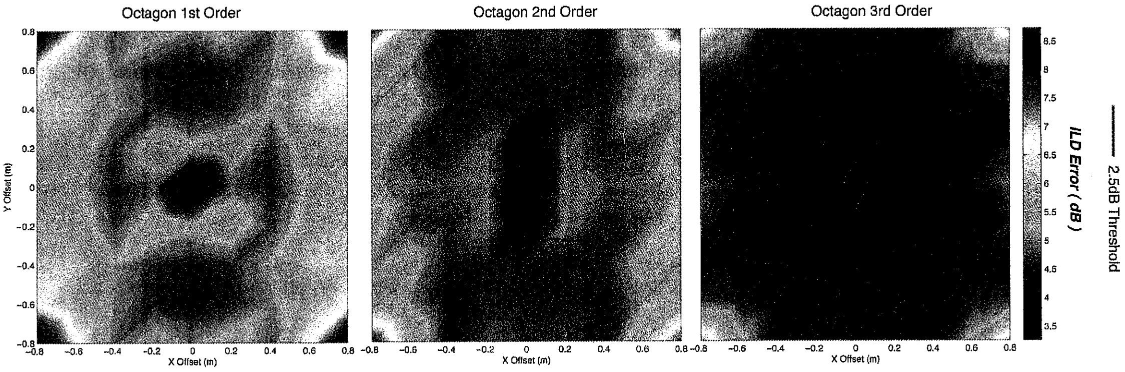

**Figure 7 - Listening area ILD error results for octagon layout**

Results highlight the following features:

1. As Ambisonic order increases, the overall level of error within the listening area is reduced.
2. Both ITD and ILD highlight reduced error in the central listening position, this behaviour is more pronounced for ITD with results highlighting clear circular patterns.
3. The 'good listening area' where ITD errors are below the perceptual threshold gets larger as Ambisonic order increases.
4. Only 3rd order ILD results show a good listening area within the threshold.

In order to try and calculate a single figure of merit for each system's ITD and ILD values, the mean average of the data was computed to compare how well Ambisonic systems perform against the ITU 5.0 layout with VBAP panning. Table 3 below presents these results up to fourth order for the octagon layout.

| System | Mean ITD Error (μs) | Mean ILD Error (dB) |
|---|---|---|
| **Ambisonic 1st Order** | 537 | 5.6 |
| **Ambisonic 2nd Order** | 460 | 5.0 |
| **Ambisonic 3rd Order** | 392 | 4.2 |
| **Ambisonic 4th Order** | 262 | 3.3 |
| **VBAP ITU 5.0** | 326 | 4.5 |

**Table 3 – Mean ITD/ILD error values for off-centre listening positions**

The results for off-centre listening positions show:

- For ITD error, 4th order Ambisonics is required to achieve an improvement over the ITU 5.0 system
- For ILD error, 3rd order Ambisonics is required to achieve an improvement over the ITU 5.0 system

## 5.1 Limitations

One of the main problems found with all ITD approximation of listening area characteristics is that time of arrival becomes more difficult to predict as the energy in an impulse response becomes spread out over time. The distance between the ears and each loudspeaker changes as a listener moves off-centre. This induced delay (between nearest and furthest loudspeaker) is around the time region in which alternative psychoacoustics mechanisms (which are also a function of relative level) start to affect perception[23], making the perceived time of arrival for the impulse response difficult to predict. While an improved method of ITD prediction may increase accuracy, a more advanced psychoacoustic model may be required to satisfactorily simulate lateralisation for multiple listening positions.

Another limitation highlighted throughout this study was the use of a spherical head model to approximate the human head. Implemented initially due to the ease of acquiring off-centre HRIRs, this model is extremely simplistic and does not account for monaural cues – an important aspect of localisation.

A further limitation worth noting is the assumption that loudspeakers act as perfect point sources for off-centre listeners. The proximity effect and changes in off-axis loudspeaker magnitude response have been identified as a significant factors in off-centre listening degradation[10], and future work could account for this behaviour.

Limited informal listening was undertaken to verify the simulation results using real speaker arrays. More formal listening tests are required in order to demonstrate the validity of the analytical binaural model and the conclusions drawn from simulated comparisons of the different reproduction systems.

# 6 CONCLUSIONS

Low-level lateralisation cues were used to measure the accuracy with which spatial audio reproduction systems re-create phantom sources. An analytical sphere model of the human head provides a good approximation to ITD and ILD cues of the KEMAR dummy head using data from the MIT database.

We conducted a simulated comparison of regular 6- and 8-speaker Ambisonic arrays with an ITU 5.0 layout that used VBAP panning. Noting that the Ambisonic and VBAP panning methods considered were tested using specific speaker layouts that differed for the two methods, the simulation results presented should not be taken as conclusive evidence that one panning method is inherently superior to another. Nevertheless the comparison of specific systems revealed some interesting results.

At the central listening position in an Ambisonic speaker array, results for ITD and ILD error in 1/3-octave frequency bands across a range of source azimuth positions were presented, indicating reduced errors as Ambisonic order increased. ITD errors for all Ambisonic systems tested were relatively low. It was shown that the 5 dB minimum ILD error threshold increased in frequency as the Ambisonic order increased which was comparable with previous studies. All Ambisonic orders tested showed improved ITD error over the VBAP ITU 5.0 layout, but 3rd order Ambisonics was needed to achieve an improved ILD error.

Off-centre listening position errors were also simulated and showed reduced ITD errors as Ambisonic order increased. Overall ILD error decreased as Ambisonic order increased but first- and second-order data was consistently above the minimum ILD error threshold. It was also highlighted that for the octagon layout tested, fourth order Ambisonics was required to improve *both* ITD and ILD error in comparison to the VBAP ITU 5.0 system.

Using low-level localisation cues to approximate off-centre listening position perception has a number of limitations. The difficulty in predicting ITD in the region of precedence effect, limitations of a spherical head model and the inability of the model to predict proximity effect and loudspeaker directivity must all be considered for future work.

# 7 REFERENCES

[1] J. Ahrens and S. Spors, "On the Scattering of Synthetic Sound Fields," in *130th Audio Engineering Society Convention*, 2011.

[2] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, Revised. London, UK: The MIT Press, 2001.

[3] V. Pulkki, “Virtual Sound Source Positioning Using Vector Base Amplitude Panning *,” *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[4] M. A. Gerzon, “Periphony : With-Height Sound Reproduction *,” *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1972.

[5] S. N. Goodwin, “3d sound for 3d games - beyond 5.1,” in *35th International Audio Engineering Society Conference*, 2009, pp. 1–8.

[6] E. Benjamin, R. Lee, and A. J. Heller, “Localization in Horizontal-Only Ambisonic Systems (revision),” in *Audio Engineering Society Convention*, 2006, pp. 1–13.

[7] G. Monro, “In-phase corrections for Ambisonics,” Sydney, Australia, 2006.

[8] D. G. Malham, “Experience With Large Area 3D Ambisonic Sound Systems,” in *Proceedings of the Institute of Acoustics 14*, 1992, pp. 209–216.

[9] B. Gardner and K. Martin, “HRTF Measurements of a KEMAR Dummy-Head Microphone,” Cambridge, MA, USA, 1994.

[10] C. Landone and M. Sandler, “Surround Sound Impact over Large Areas,” in *110th Audio Engineering Society Convention*, 2001.

[11] J. S. Asvestas, J. J. Bowman, P. L. Christiansen, O. Einarsson, R. E. Kleinman, D. L. Sengupta, T. B. A. Senior, F. B. Sleator, P. L. E. Uslenghi, and N. R. Zitro, *Electromagnetic and Acoustic Scattering by Simple Shapes (Revised Printing)*. New York: Hemisphere Publishing Corporation, 1987.

[12] R. V. L. Hartley and T. C. Fry, “The Binaural Location of Pure Tones,” *Phys. Rev.*, vol. 18, no. 6, pp. 431–442, 1921.

[13] L. A. Jeffress, “A place theory of sound localization,” *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[14] D. J. Kistler and F. L. Wightman, “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction.,” *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637–47, Mar. 1992.

[15] J. Estrella, “On the Extraction of Interaural Time Differences from Binaural Room Impulse Responses,” TU Berlin, Germany, 2010.

[16] E. Bates, G. Kearney, and D. Furlong, “Localization accuracy of advanced spatialisation techniques in small concert halls,” *Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 3069–3070, 2007.

[17] S. Busson, R. Nicol, and B. F. G. Katz, “Subjective investigations of the interaural time difference in the horizontal plane,” *in 118th Audio Engineering Society Convention*, preprint 6324, 2005.

[18] W. Gaik, “Combined evaluation of interaural time and intensity differences: psychoacoustic results and computer modeling.,” *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 98–110, Jul. 1993.

[19] V. Pulkki and T. Hirvonen, “Localization of Virtual Sources in Multichannel Audio Reproduction,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 105–119, 2005.

[20] J. Daniel, J. Rault, and J. Polack, “Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions,” in *105th Audio Engineering Society Convention*, 1998, vol. 4795.

[21] R. M. Aarts, “A Comparison of Some Loudness Measures for Loudspeaker Listening Tests,” *Journal of the Audio Engineering Society*, vol. 40, no. 3, 1992.

[22] D. Moore and J. Wakefield, “Exploiting Human Spatial Resolution in Surround Sound Decoder Design,” in *125th Audio Engineering Society Convention*, 2008.

[23] F. Rumsey, *Spatial Audio*, 2nd Editio. Oxford, UK: Focal Press, 2001.