# The analysis of multi-channel sound reproduction algorithms using HRTF data

B. Wiggins, I. Paterson-Stephens, P. Schillebeeckx
Signal Processing Applications Research Group
University of Derby
Derby, United Kingdom

Described in this paper is a method for the analysis and comparison of multi-speaker surround sound algorithms using HRTF data. Using Matlab and Simulink [1] a number of surround sound systems were modeled, both over multiple speakers (for listening tests) and using the MIT Media Labs HRTF set (for analysis)[2]. The systems under test were 1st Order Ambisonics over eight and five speakers, 2nd Order Ambisonics over eight speakers and Amplitude panned 5.0 over five speakers. The listening test results were then compared to the HRTF analysis with favourable results.

## INTRODUCTION

Much research has been carried out in to the performance of multi-channel sound reproduction algorithms, both subjectively and objectively. Much of the quantitative data available on the subject has been calculated by mathematically simulating acoustical waves emitting from a number of fixed sources (speakers) [3,4]. The resulting sound field can then be observed. This method, although giving a good overview of the systems performance in a space, does not lend itself well to an analysis of how well a subject can localise sound sources using a particular system. In this paper, a method of analysis will be described using head related transfer functions as a reference for the localisation cues needed to successfully localise a sound in space. This method will then be compared to results obtained from a recent listening test carried out at the University of Derby's Multi Channel Sound Research Laboratory.

## ANALYSIS USING HRTF DATA

The underlying theory behind this method of analysis is that of simple comparison. If a real source travels through $360^0$ around the head (horizontally) and the sound pressure level at both ears is recorded, then the three widely accepted psycho-acoustic localisation cues [5,6] can be observed: The time difference between the sounds arriving at each ear due to different path lengths, the level difference between the sounds arriving at each ear due to different path lengths and head shadowing, and pinna filtering, a combination of complex level and time differences due to the listeners own pinna. The most accurate way to analyse and/or reproduce these cues is with the use of head related transfer functions.

For the purpose of this analysis technique, the binaural synthesis of virtual sound sources is taken as the reference system as the impulse responses used for this system are of real sources in real locations. The HRTF set used do not necessarily need to be optimal for all listeners (which can be an issue for binaural listening) so long as all of the various localisation cues can be easily identified. This is the case because this form of analysis compares the *difference* between real and virtual sources and as all systems will be synthesised using the same set of HRTFs, there performance next to another set should not be of great importance.

Once the system has been synthesised using HRTFs, impulse responses can be calculated for virtual sources from any angle so long as the panning laws for the system to be tested are known. Once these impulse responses have been created the three parameters used for localisation can be viewed and compared, with estimations made as to how well a particular system is able to produce accurate virtual images.

### Advantages Of This Technique

- *All* forms of multi-channel sound can potentially be analysed meaningfully using this technique.
- Direct comparisons can be made between very different multi-channel systems as long as the HRTFs used to analyse the systems are the same.
- Systems can be auditioned over headphones.

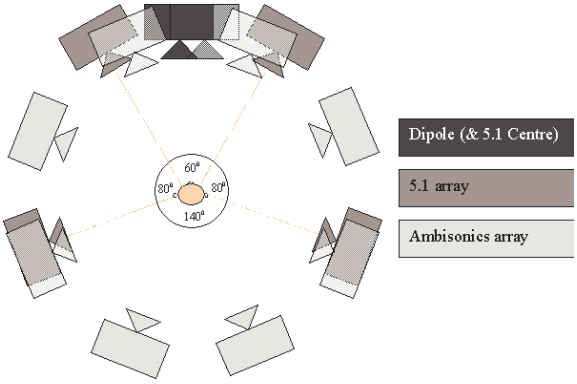
## LISTENING TESTS

In order to have a set of results to use as a comparison for this form of analysis a listening test was carried out. The listening test comprised of a set

of ten tests for five different forms of surround sound:

- 1st Order Ambisonics over 8 speakers (horizontal only)
- 2nd Order Ambisonics over 8 speakers (horizontal only)
- 1st Order Ambisonics over a standard 5 speaker layout.
- Amplitude panned over a standard 5 speaker layout.
- Stereo Dipole using two speakers at +/- $5^0$.

The tests were to be carried out in the University of Derby's Multi Channel Sound Research Laboratory with speakers setup as shown in figure 1.

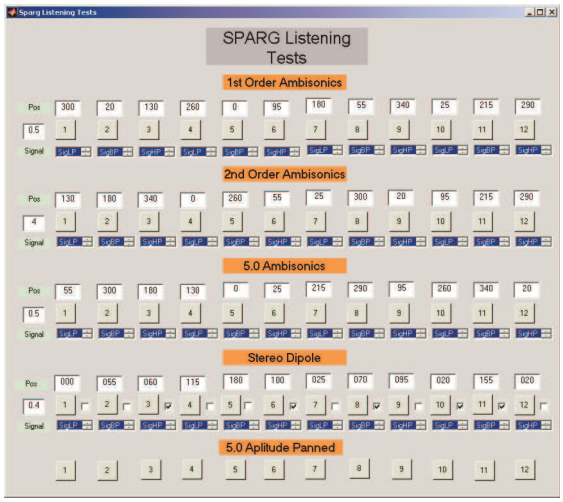



[Figure 1] Layout of Multi-channel Sound Research Lab.

The listing room has been acoustically treated and a measurement of the ambient noise in the room gave around 43dBA in most $^1/_3$-octave bands, with a peak at 100Hz of 52.1dBA and a small peak at 8kHz of 44.4dBA. The RT60 of the room is 0.42 seconds on average, but is shown in $^1/_3$-octave bands in figure 15.

Using a PC and a multi-channel soundcard (Soundscape Mixtreme) all of the speakers could be accessed simultaneously [1], if needed, and so tests on all of the systems could be carried in a single session.

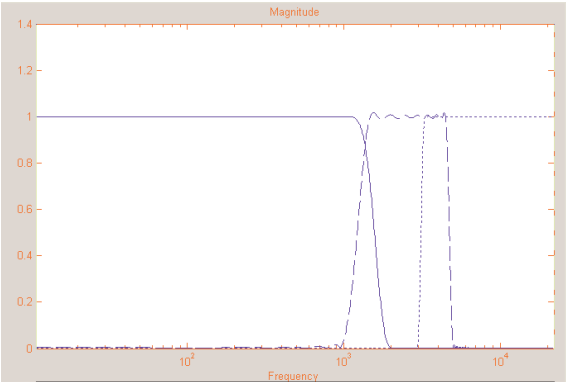
A flexible framework was devised using Matlab and Simulink (The Mathworks, Inc) so that listening test variables could be changed with minimal effort, with the added bonus that the framework would be reusable for future tests. A Simulink 'template' file was created for each of the five systems that could take variables from the Matlab workspace, such as input signal, overall gain and panning angle. Then a GUI was created where all of the variables could be entered and the individual tests run. A screen shot of the final GUI is shown in figure 2.




[Figure 2] Screen shot of listening test GUI.

The overall gain parameter was included so each of the different systems could be configured to have a similar subjective gain, with the angle of the virtual source specified in degrees. The only exception to this was the 5.0 Amplitude panned system where the speaker feeds were calculated off line using the Mixtreme soundcards internal mixing feature. The amplitude panning algorithms will be included in the next version of the GUI. Also, the extra parameter (tick box) in the stereo dipole section was used to indicate which side of the listener the virtual source would be placed as the HRTF set used [2] only had impulse responses for the right hemisphere and must be reversed in order to simulate sounds originating from the left (indicated by a tick).

There were three separate sources used in this test. These signals were band limited pulsed noise, three pulses per signal, with each pulse lasting two seconds with one second of silence between each pulse. Each signal was band limited according to one of the three localisation frequency ranges taken from two texts [5,6]. These frequencies are not to be taken as absolutes, just a starting point for this line of research. A plot of the frequency ranges for each of the three signals is shown in figure 3.




[Figure 3] Filters used for listening test signals.

Twenty eight test subjects were used, most of whom had never taken part in a listening test before. The test subjects were all enrolled on the 3rd year of the University's Music Technology and Audio System Design course, and so knew the theory behind some surround sound systems, but had little or no listening experience of the systems at this point. Each listener was asked to try to move their head as little as possible while listening, and to indicate the direction of the source by writing the angle, in degrees, on an answer paper provided. Listeners could ask to hear a signal again if they needed to, and the operator only started the next signal after an answer had been recorded. The listeners were also given a sheet of paper to help them with angle locations with all of the speaker positions marked in a similar fashion to figure 1, except without the surround sound system labels, and with $0^0$ to $360^0$ marked out in $5^0$ intervals.

## HRTF SIMULATION

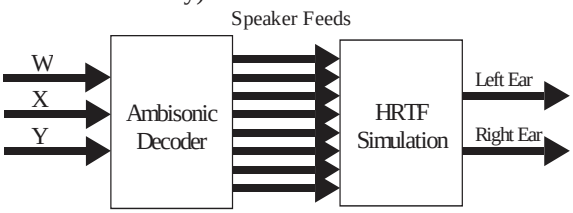
For the scope of this paper three of the five systems will be analysed using the HRTF method described above:

- 1st Order Ambisonics
- 2nd Order Ambisonics
- 1st Order Ambisonics over 5 speakers.

The listening test results for the amplitude panned 5 speaker system will also be included, however.

The set of HRTFs used for this analysis were the MIT media lab set of HRTFs, specifically the compact set [2]. As mentioned earlier, it is not necessarily important that these are not the best HRTF set available, just that all of the localisation cues are easily identifiable.

All systems can be simulated binaurally but Ambisonics is a slightly special case as it is a matrixed system comprising of the steps shown in figure 4 (see references [1,3,4,7,8] for discussions on Ambisonics theory).




[Figure 4] Block diagram of the Ambisonic to binaural conversion process.

Because the system takes in three channels which are decoded to eight speaker feeds, which are then decoded again, to two channels, the intermediate decoding to eight speakers can be incorporated into the HRTFs calculated for W, X and Y meaning that only six individual HRTFs are needed for *any* speaker arrangement [Equ. 1]. If the head is assumed to be symmetrical (which they are in the MIT set of compact HRTFs) then even less HRTFs are needed as $W_{left}$ and $W_{right}$ will be the same (Ambisonics omni-directional component), $X_{left}$ and $X_{right}$ will be the same (Ambisonics front/back component) and $Y_{left}$ will be $180^0$ out of phase with respect to $Y_{right}$. This means a whole 1st order Ambisonic system comprising of any amount of speakers can be simulated using just three HRTF filters.

$$W^{hrtf} = \left(\sqrt{2}\right) \times \sum_{k=1}^{8} \left(S_k^{hrtf}\right)$$

$$X^{hrtf} = \sum_{k=1}^{8} \left(\cos(\theta_k)\sin(\phi_k) \times S_k^{hrtf}\right)$$

$$Y^{hrtf} = \sum_{k=1}^{8} \left(\sin(\theta_k)\sin(\phi_k) \times S_k^{hrtf}\right)$$

Where
$\theta$ = source azimuth
$\phi$ = source elevation (0 for horizontal only)
$S_k^{hrtf}$ = Pair of Speakers positional HRTFs.

[Equ 1] 1st Order Ambisonics to binaural conversion.

Once the HRTFs for W, X and Y are known a virtual source can be simulated by using the first order Ambisonics encoding equations shown in Equ. 2 [7].

$$W = \left(1/\sqrt{2}\right) \times x(n)$$

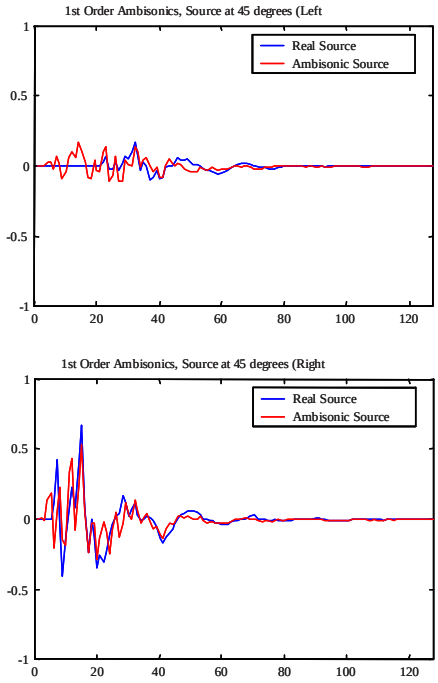
$$X = \cos(\theta) \times \sin(\phi) \times x(n)$$

$$Y = \sin(\theta) \times \sin(\phi) \times x(n)$$

Where x(n) is the signal to be placed in virtual space.

[Equ 2] 1st Order Ambisonics encoding equations

Using two sets of the W, X and Y HRTFs (one for eight and one for five speaker 1st order Ambisonics) and one set of W, X, Y, U and V [4,9] for the 2nd order Ambisonics, sources were simulated from 0 to $360^0$ in $5^0$ intervals. The $5^0$ interval was dictated by the HRTF set used, as although the speaker systems could now be simulated for any source angle, the real sources (used for comparison) could only be simulated at $5^0$ intervals (without the need for interpolation). An example pair of HRTFs for a real and a virtual source are shown in figure 5.




[Figure 5] Example left and right HRTFs for a real and virtual source (1st Order Ambisonics) at $45^0$ anticlockwise from centre front.

## Impulse Response Analysis

As mentioned in the introduction, three localisation cues will be analysed, interaural level difference, interaural time difference, and pinna filtering effects. The impulse responses contain all three of these cues together meaning that although a clear filter delay and level difference can be seen by inspection, the pinna filtering will make both the time and level differences frequency dependant. These three cues will be calculated using the following methods:

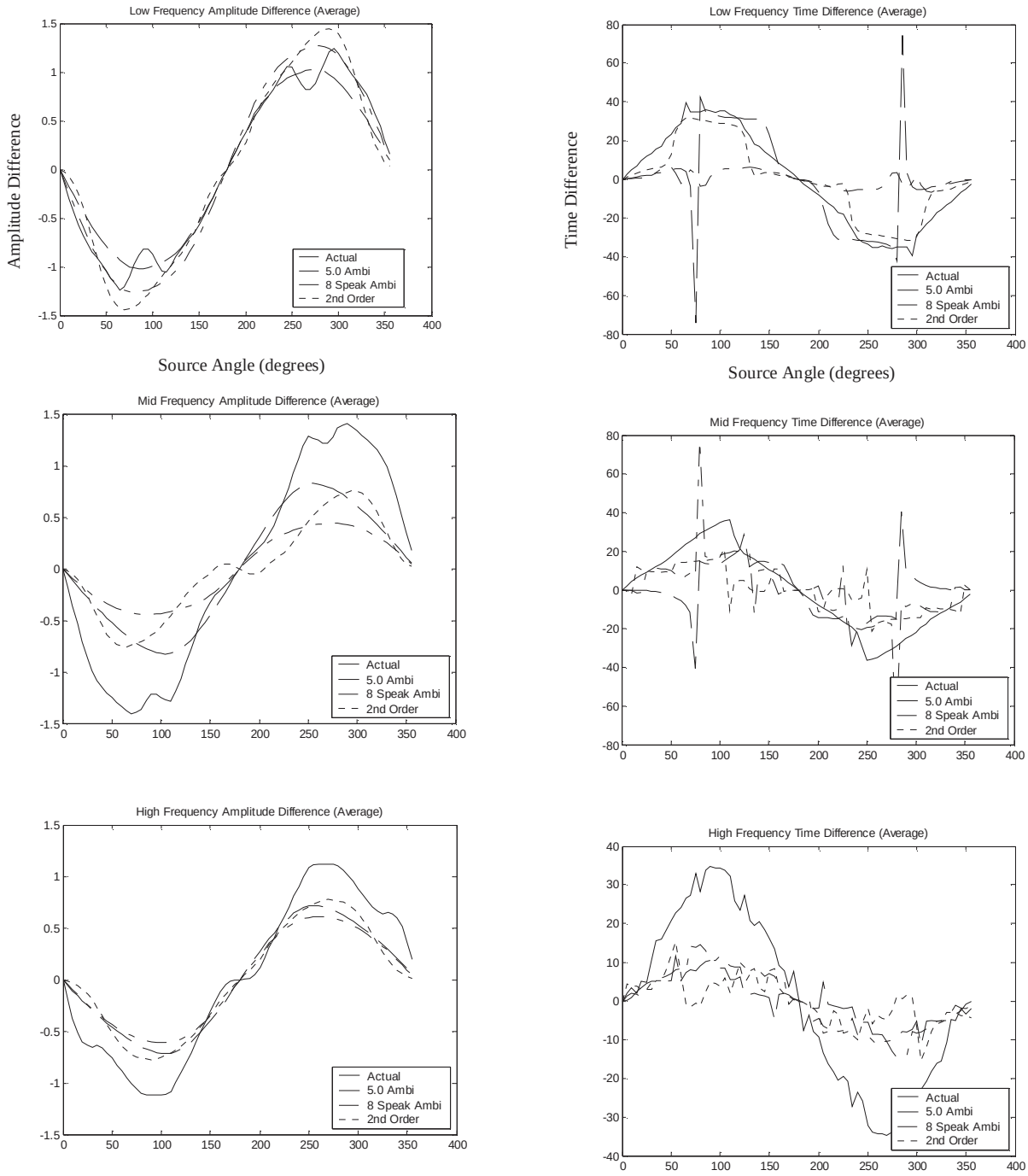
- Interaural Amplitude Difference – Mean amplitude difference between the two ears, taken from an FFT of the impulse responses.
- Interaural Time Difference – Mean time difference between the two ears, taken from the group delay of the impulse responses.
- Pinna filtering – Actual time and amplitude values, taken from the group delay and an FFT of the impulse responses.

Once the various psycho-acoustic cues have been separated, comparisons can be made with the cues of an actual source and estimations of where the sounds may appear to come from can be made using each of the localisation parameters in turn. As the analysis is carried out in the frequency domain, band limiting the results (to coincide with the source material used in the listening tests) is just a case of ignoring any data that is outside the range to be tested.

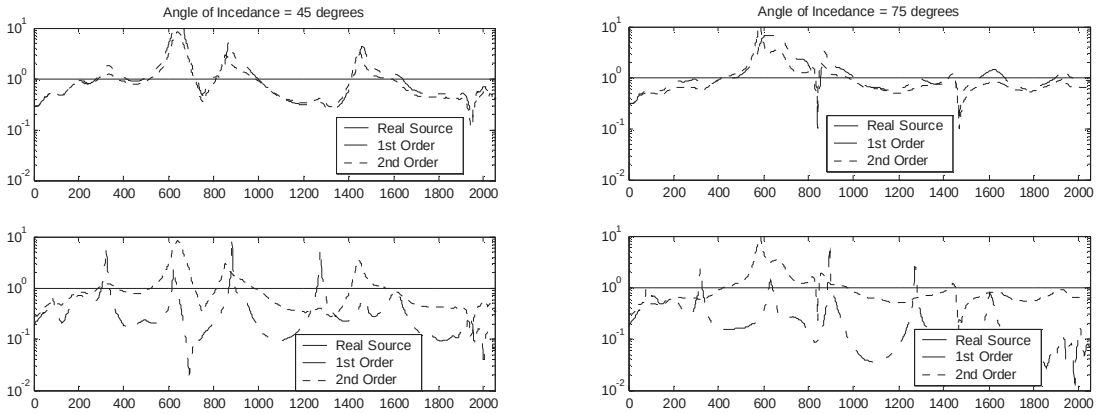
As an example, figure 6 shows the low, mid and high frequency results for real sources and the three Ambisonic systems for averaged time and amplitude differences between the ears.

These graphs show a number of interesting points about the various Ambisonic systems. Firstly, the 2nd order system actually has a *greater* amplitude difference between the ears at low frequencies when compared to a real source, and this is also the frequency range where all of the systems seem to correlate best with real sources. However, the ear tends to use amplitude cues more in the mid frequency range, and another unexpected result was also discovered here. It seems that the 1st order, five speaker system actually outperforms the 1st order, eight speaker system at mid frequencies, and seems to be equally as good as the eight speaker, second order system. This is not evident in the listening tests, but if the average time difference graphs are observed it can be seen that the five speaker system has a number of major errors around the $90^0$ and $270^0$ source positions and shows the 2nd order system to hold the best correlation. The time difference plots all show that the five speaker system still outperforms the 1st order, eight speaker system, apart from the major disparities, mentioned above, at low frequencies. It can be seen from the listening test results (figure 11) that the five speaker system does seem to be at least as good as the eight speaker system in all three of the frequency ranges, which was not expected. The mid and high frequency range graphs are a little too complicated to analyse by inspection and so will be looked at later in the paper using a different technique.

The pinna filtering can also be clearly seen in the simulation, but is a more complex attribute to analyse




[Figure 6] Graphs to show the average amplitude and time differences between the ears for low, mid and high frequency ranges.




[Figure 7] Graphs to show the difference in pinna amplitude filtering of a real source and 1st and 2nd order Ambisonics (eight speaker) when compared to a real source.

directly, although it has been useful to look at for a number of reasons. If the amplitude or group delay parameters are looked at over the full $360^0$ it can be seen that they both change radically due to virtual source position (as does a source in reality). However, the virtual sources change differently when compared to real sources. This change will also occur if the head is rotated (in the same way for a regular rig, or a slightly more complex way for an irregular five speaker set-up) and I believe that this is the 'phasiness' that Gerzon often mentioned in his papers regarding the problems of Ambisonics [3]. This problem, however, is not strictly apparent as a timbral change when a source or the listeners' head moves, but instead probably just aids in confusing the brain as to the sound sources real location, increasing source location ambiguity and source movement when the listeners head is turned. This parameter is more easily observed using an animated graph, but is shown as a number of stills in figure 7.

Due to the complexity of the results obtained using the HRTF simulation for the pinna filtering, it is difficult to utilise these results in any estimation of localisation error, although further work will be carried out to make use of this information. However, using the average time and amplitude differences to estimate the perceived direction of the virtual sound source is a relatively trivial task using simple correlation between the actual and virtual sources. Figures 8,9 and 10 show the listening test results with the estimated localisations also shown, using the average amplitude and the average time differences at low and mid frequencies.

The listening tests themselves, gave reasonably expected results as far as to the system that performed best (the 2nd Order Ambisonics system). However the other three systems (1st order eight and five speaker, and amplitude panned 5.0) all seemed to perform equally as well, which was not expected. This may have been because the five speaker set up consisted of better quality speakers than the eight speaker rig. The frequency content of the sounds did not seem to make any difference in the perceived localisation of the sound sources, although a more extensive test would have to be undertaken to confirm this, as the purpose of this test was just to see if there were any major differences between the three localisation frequency ranges. Another interesting result was the virtual source at $0^0$ on the amplitude panned system (see figure 11). As there is a centre front speaker, a virtual source at $0^0$ just radiates from the centre speaker, i.e. it is a *real* source at $0^0$. However, around 30% of the subjects recorded that the source came from behind them. Front/back reversals were actually less common in all of the other systems (at $0^0$), apart from 2nd order Ambisonics (the system that performed best).

The source position estimation gave reasonably good results when compared with the results taken from the listening tests, with any trends above or below the diagonal, representing a perfect score, being estimated successfully. If the graphs represented truly what is expected from the different types of psycho-acoustic sound localisation, then the low frequency time graph and the mid frequency amplitude graph should make the best indicator as to where the source is coming from. However it is well known [5] that if one localisation cue points to one direction, and the other cue points to another, then it may be some direction between these two localisation angles that the sound is actually perceived to originate from. The HRTF analysis does not take this into account at the moment and so some error is

expected. Also, the compact set of HRTFs used are the minimum phase versions of the actual HRTFs recorded which may contribute to the time difference estimation results (although the cues seem reasonable when looked at for the actual sources). As mentioned, there was no major difference between the three different signals in terms of localisation error. Because of this the plots showing the estimated localisation using the whole frequency range are shown in figures 12-14 which also show the interaural amplitude difference as a better localisation approximation.

## CONCLUSIONS

The HRTF analysis of the three surround systems described in this paper seems to work reasonably well, even at this early stage, and the method is definitely worth perusing as a technique that can be used to evaluate and compare *all* forms of surround sound systems equally. Although the errors seen in the estimation when compared to the listening test results can be quite large, the general trends were shown accurately, even with such a simple correlation model used.

## FURTHER WORK

More extensive listing tests need to be carried out in order to generate results for a greater number of source positions and subjects, so a more obvious average perceived localisation can be used as a comparison. The source material must also be reviewed with the overlapping frequency ranges being changed so that more of a difference between them is apparent by perhaps using more frequency ranges. Different sets of HRTFs will also be tried, although this is not expected to affect the results significantly as the analysis works on comparisons using the actual HRTF data as a reference.

## REFERENCES

[1] Schillebeeckx P., Paterson-Stephens I., Wiggins B., "Using Matlab/Simulink as an implementation tool for Multi-Channel Surround Sound", Proceedings of the 19th International AES conference on Surround Sound, June 2001.

[2] Gardner B., Martin K., "HRTF Measurements of a KEMAR Dummy-Head Microphone", 1994. http://sound.media.mit.edu/KEMAR.html

[3] Gerzon M., "Psychoacoustic Decoders for Multispeaker Stereo and Surround Sound", Proceedings of the 93rd AES Convention, October 1992.

[4] Bamford J., "An Analysis of Ambisonic Sound Systems of First and Second Order", Thesis submitted to the University of Waterloo, Ontario, Canada, 1995.

[5] Gulick W., Gescheider G., Frisina R., "Hearing – Physiological Acoustics, Neural Coding and Psychoacoustics", Chapter113, Oxford Press 1989.

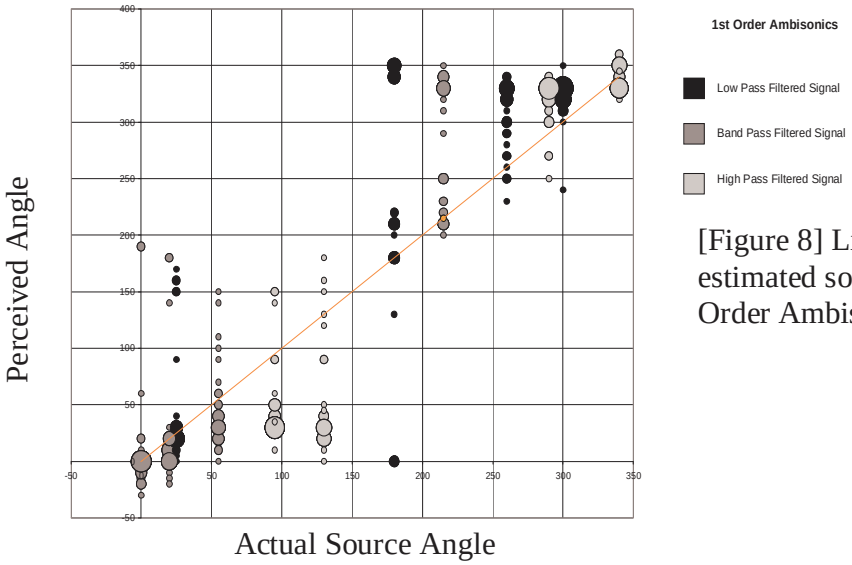
[6] Rossing T., "The Science of Sound", Chapter 5, Addison Wesley, 1990.

[7] Malham D., "Spatial hearing mechanisms and sound reproduction", University of York, 1998. http://www.york.ac.uk/inst/mustech/3d_audio/ambis2.htm

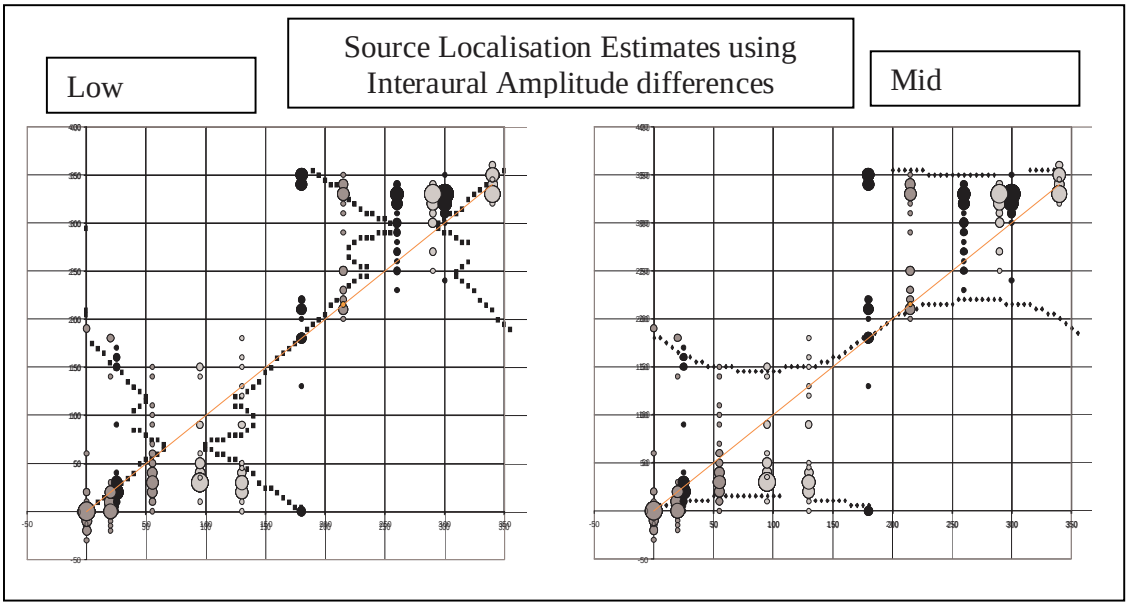
[8] Farino A., Ugolotti E., "Software Implementation of B-Format Encoding and Decoding", Proceedings of the 104th AES Convention, May 1998.

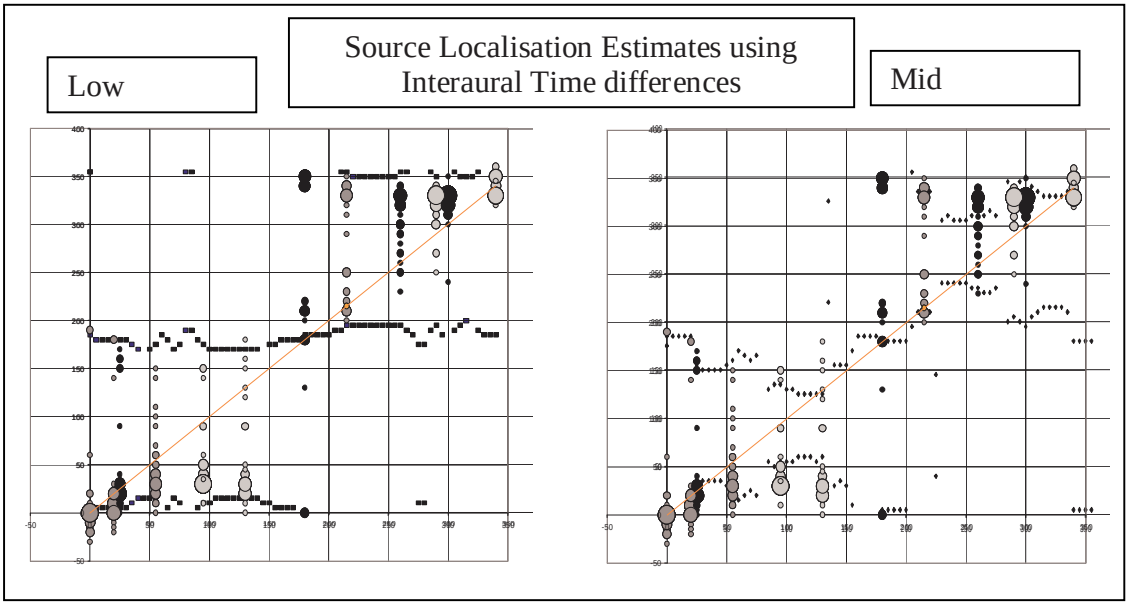
[9] Furse R., "3D Audio Links and Information", http://www.muse.demon.co.uk/3daudio.html ,

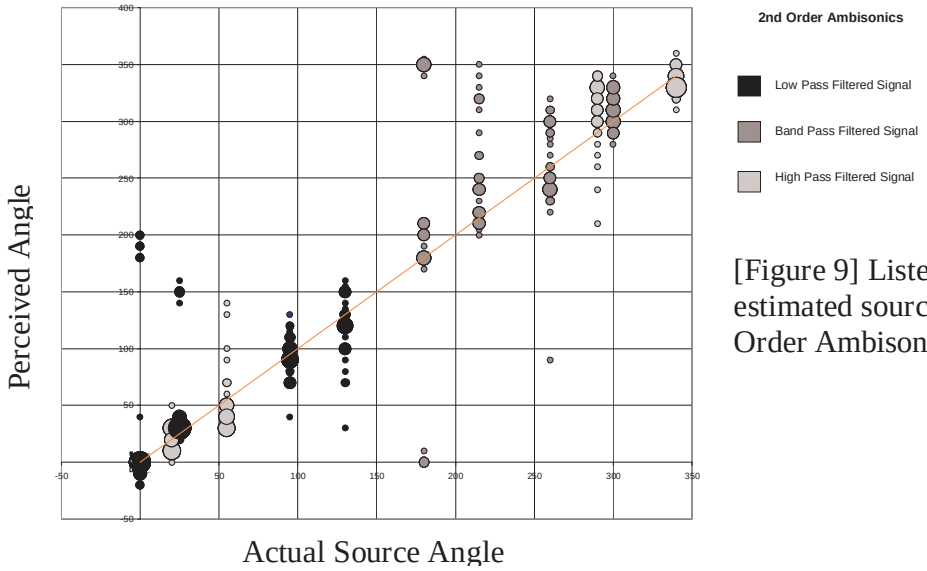
[10] Paterson-Stephens I., Bateman A., "The DSP Handbook, Algorithms, Applications and Design Techniques", Prentice Hall, 2001.



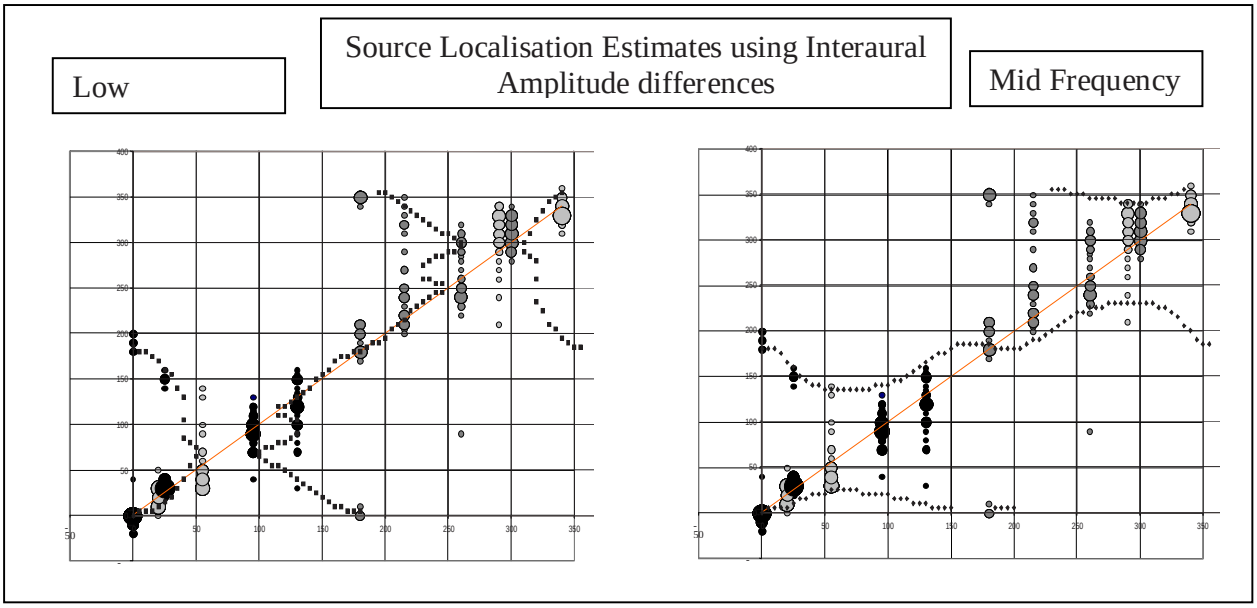

[Figure 8] Listening Test results and estimated source localisation for 1st Order Ambisonics

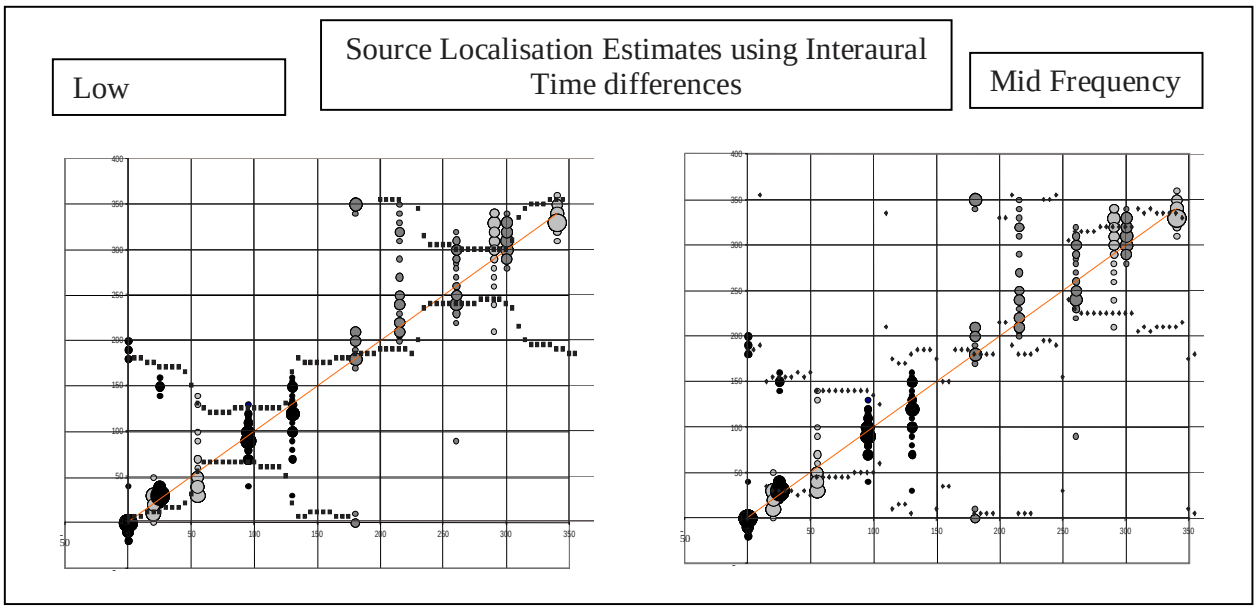


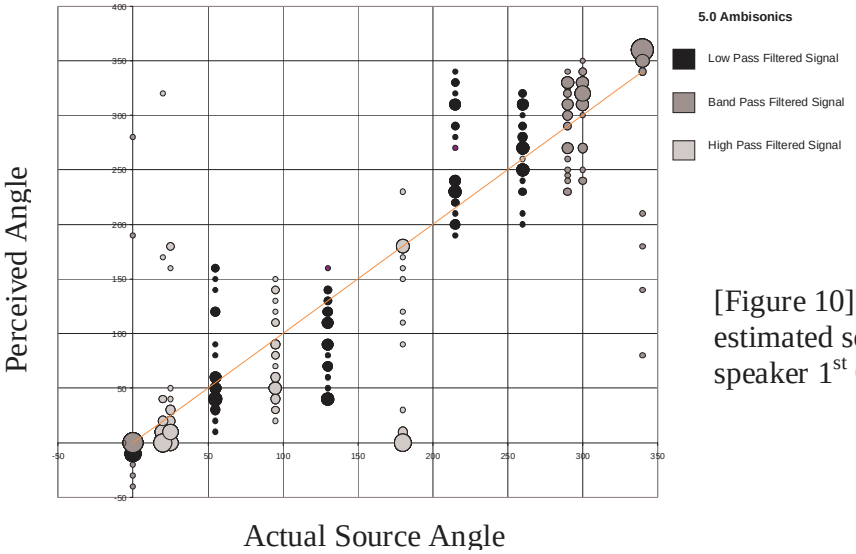





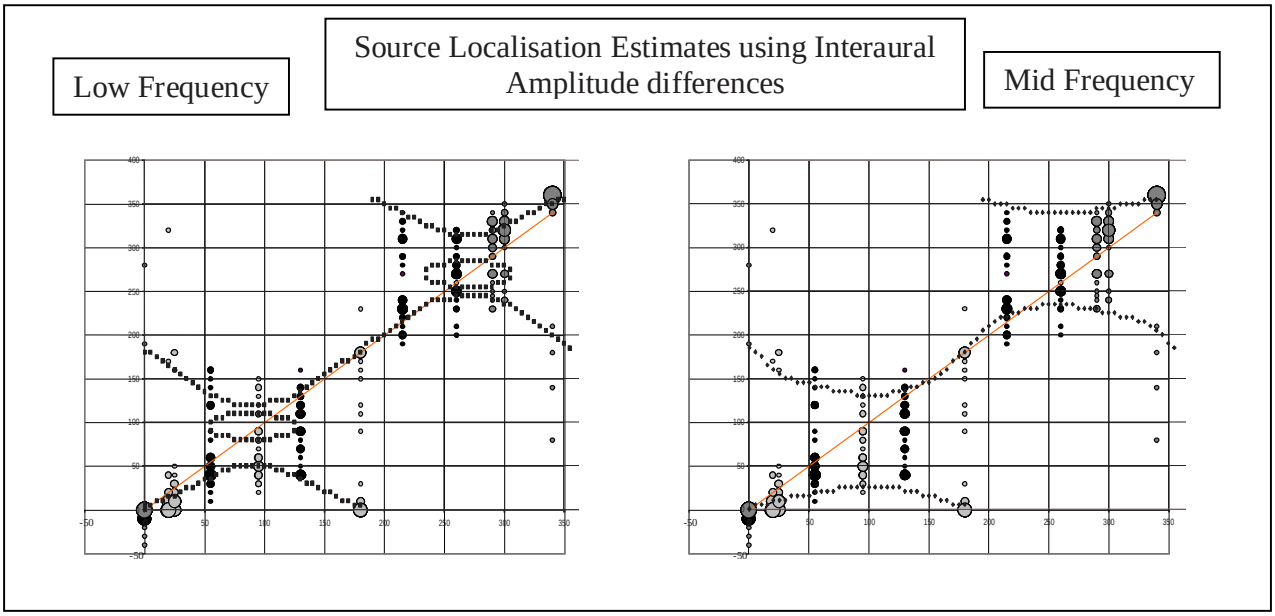

[Figure 9] Listening Test results and estimated source localisation for 2nd Order Ambisonics

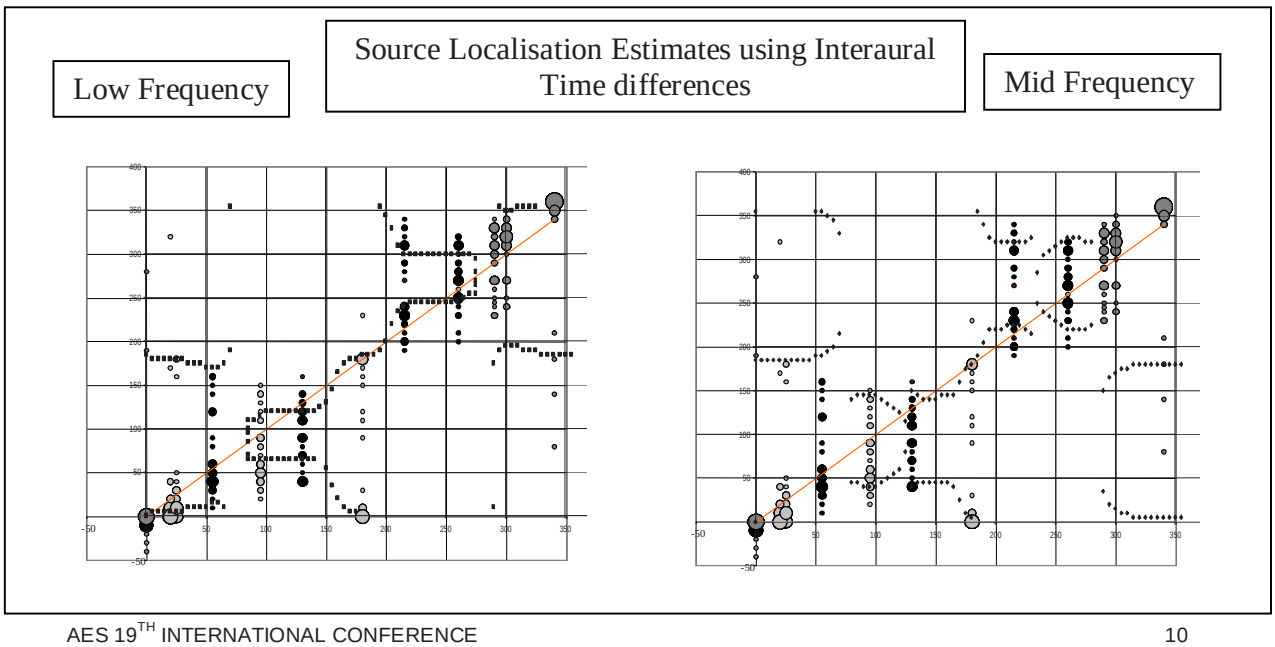


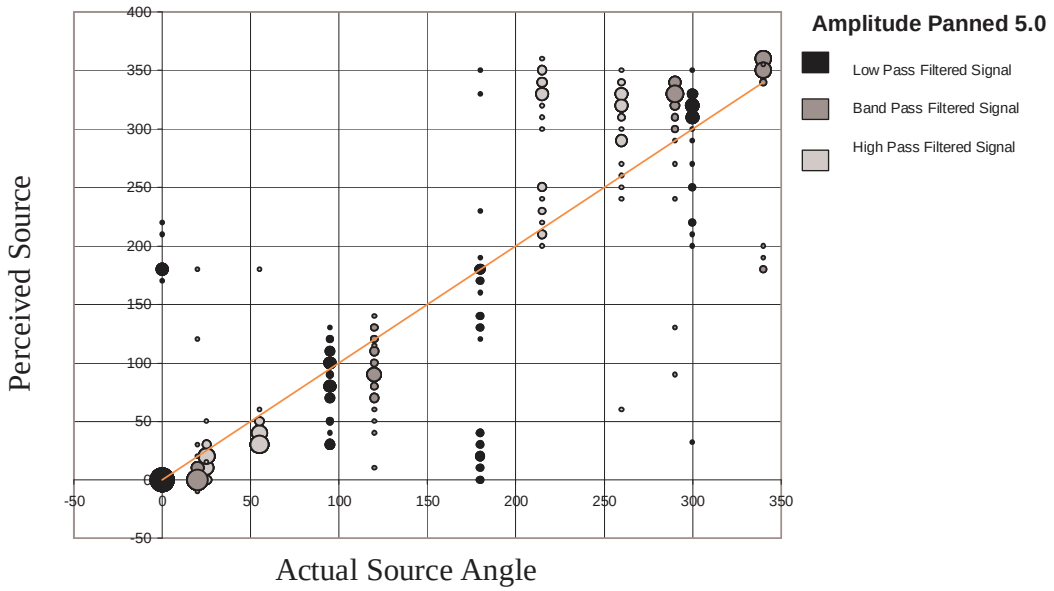





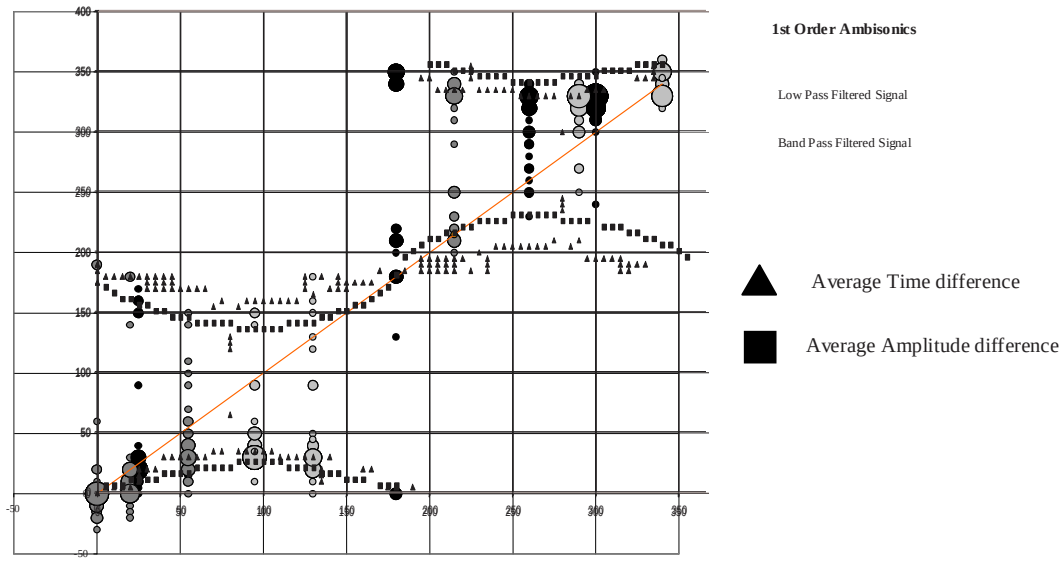

[Figure 10] Listening Test results and estimated source localisation for five speaker 1st Order Ambisonics



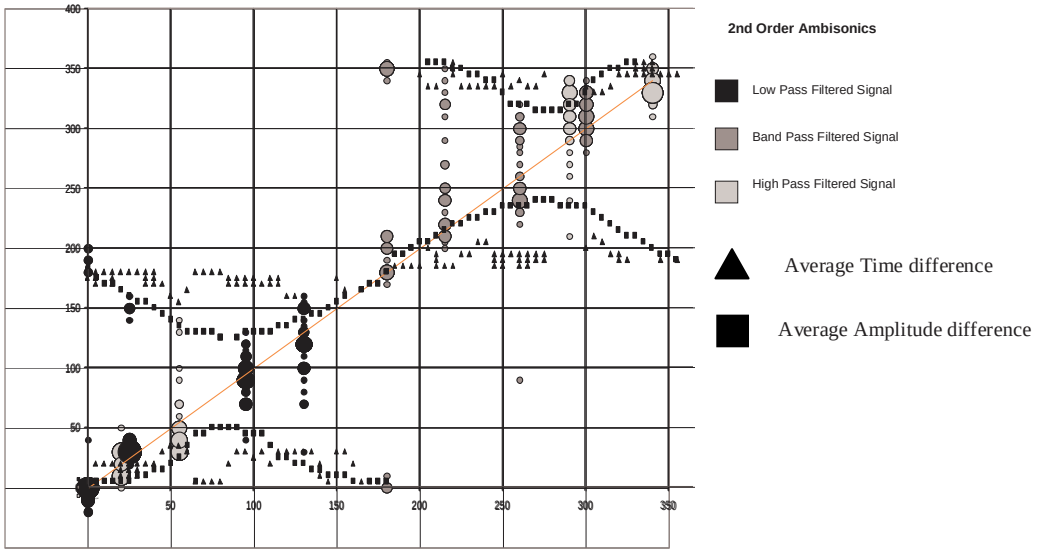





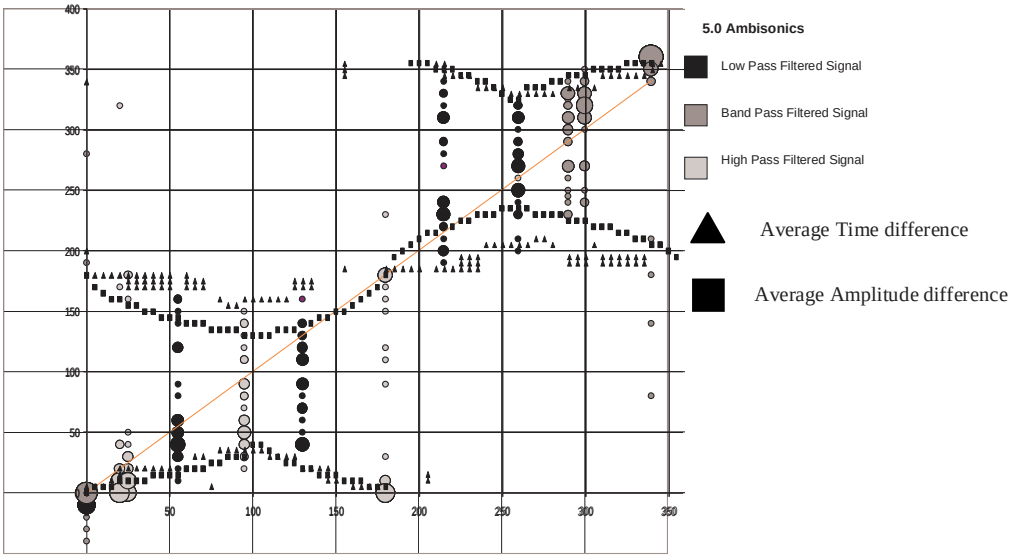

[Figure 11] Listening test results for Amplitude Panned five speaker system.



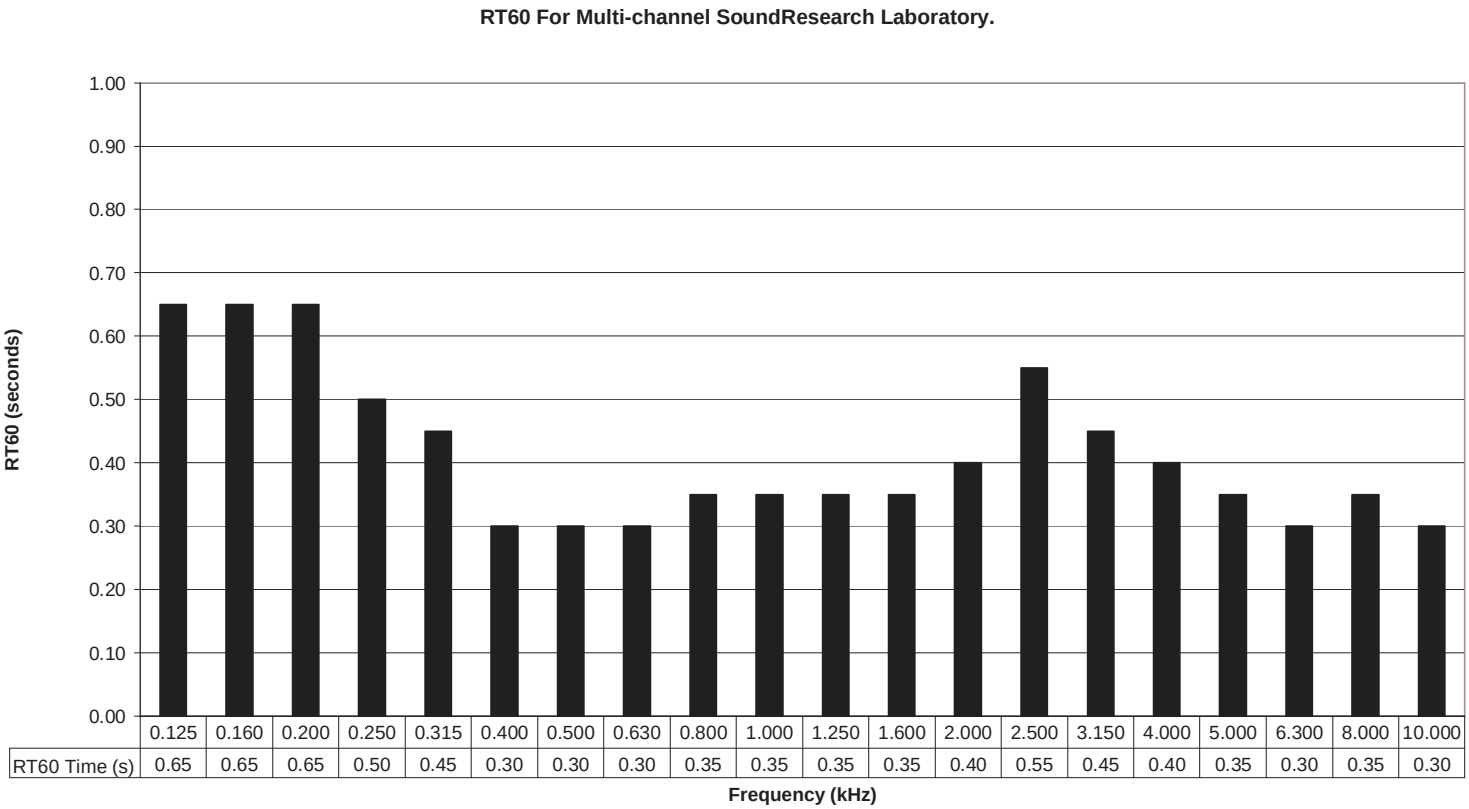

[Figure 12] Average Time and Frequency Localisation Estimate for 1st Order Ambisonics.




[Figure 13] Average Time and Frequency Localisation Estimate for 2nd Order Ambisonics.




[Figure 14] Average Time and Frequency Localisation Estimate for five speaker 1st Order Ambisonics.




[Figure 15] RT60 Measurement of the University of Derby's multi-channel sound research laboratory, shown in $^{1}/_{3}$ octave bands.