Audio Engineering Society
Convention Paper
Presented at the 110th Convention
2001 May 12–15 Amsterdam, The Netherlands

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

An investigation of interaural time difference fluctuations, part 2: dependence of the subjective effect on audio frequency

Russell Mason, Francis Rumsey and Bart de Bruyn,
Institute of Sound Recording and Department of Psychology,
University of Surrey,
Guildford, Surrey. GU2 7XH, UK

ABSTRACT

The effect of the audio frequency of narrow-band noise signals with a sinusoidal ITD fluctuation was investigated. To examine this, a subjective experiment was carried out using a match to sample method and stimuli delivered over headphones. It was found that the magnitude of the subjective effect is dependent on audio frequency and that the relationship between the audio frequency and a constant subjective effect appears to be based on equal maximum phase difference fluctuations.

INTRODUCTION

Interaural time difference (ITD) fluctuations are changes over time of the relative phase between the two audio signals measured at the ears. If the relative phase fluctuates slowly, the subjective effect is a change in the perceived position of a sound. However, if the fluctuations occur at a frequency above a few Hertz, the perception is no longer movement due to the perceptual effect of 'binaural sluggishness' or 'localisation lag' [1, 2]. Grantham and Wightman researched this effect, and found that ITD fluctuations at a rate of greater than approximately 20 Hz caused the perception of width or diffuseness instead of movement [3]. Griesinger described the subjective effect as a 'stationary source in the presence of a surround' [2].

These ITD fluctuations are created in real acoustic environments, caused by the interaction of a direct source signal with the reflections from a number of objects or boundaries within that acoustic environment [4]. Therefore the magnitude of these fluctuations could be calculated in order to create an acoustic measurement that correlates with the subjective perception of acoustic environments, either real or reproduced.

Two measurement techniques that measure these ITD fluctuations have recently been proposed. The first, the diffuse field transfer function (DFT), measures the difference between zero-crossing points in the audio signals reaching each ear to derive the ITD fluctuations [5]. The second, the interaural cross-correlation fluctuation function (IACCFF), uses a series of interaural cross-correlation measurements to achieve the same result [6].

Research into ITD fluctuations has shown there to be a lack of understanding of the phenomenon. In order to refine these measurements, a number of experiments need to be undertaken. This series of papers documents some of the work that has been carried out to understand the subjective effect of the ITD fluctuations more comprehensively, and to answer some specific questions related to the measurement of these fluctuations.

This paper concerns the measurement of ITD fluctuations. It is apparent that the dependence of the subjective effect of the ITD fluctuations on the audio frequency of the sound has not been investigated. It may be that a narrow-band audio signal at a low audio frequency with a given time magnitude[1] of sinusoidal ITD fluctuations has subjectively the same spatial effect as a narrow-band audio signal at a higher audio frequency with a much lower time magnitude of sinusoidal ITD fluctuations. If this is the case and it is not taken into account in the measurement technique then it may have a large effect on the accuracy of a measurement, especially if the source signals are narrow-band. Therefore this factor needs to be investigated.

It must be noted that the terminology used in this paper for the subjective spatial effect created by these ITD fluctuations is deliberately vague. This is because a set of descriptors for the subjective effect has not been developed sufficiently. The previous experiment reported in [7] attempted to elicit the subjective effect of continuous noise signals with different sinusoidal ITD fluctuations delivered over headphones. It uncovered that there were inter-subject differences in the responses and so the subjective spatial effect could not be defined accurately. Because of this uncertainty, the subjective effect of the noise signals with specific sinusoidal ITD fluctuations will be described in this paper as the 'subjective spatial effect'.

AIMS

The aims of the experiment were twofold. Firstly, to find whether the magnitude of the subjective spatial effect created by audio signals with specific interaural time difference (ITD) fluctuations are dependent on the audio frequency. Secondly, to investigate the relationship between a wide-band sample with a given phase magnitude of sinusoidal ITD fluctuations and a narrow-band segment taken from the wide-band sample.

To explain these aims in more detail, the possible outcomes of the experiment are explored.

Aim 1: dependency of magnitude of subjective effect on audio frequency

The first aim of the experiment was to find whether the subjective spatial effect of two narrow-band stimuli, with the same ITD fluctuation frequencies and time magnitudes but different audio frequencies, were perceived to have a similar magnitude of subjective spatial properties or not. The time magnitude of sinusoidal ITD fluctuations can be visualised by plotting a graph of audio frequency against maximum ITD. The latter is the peak ITD of the sinusoidal fluctuation, measured in time from the zero position[2].

The plots in Figure 1 show narrow-band audio samples centred on two different audio frequencies, with the audio samples at the higher frequency having two different time magnitudes of fluctuation.

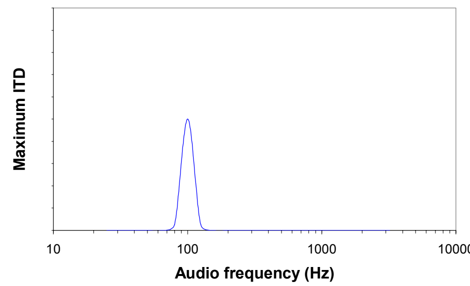


Figure 1a: Plot of audio frequency against maximum interaural time difference (ITD) of a narrow-band sample centred on 100 Hz with a sinusoidal ITD fluctuation with a large maximum ITD

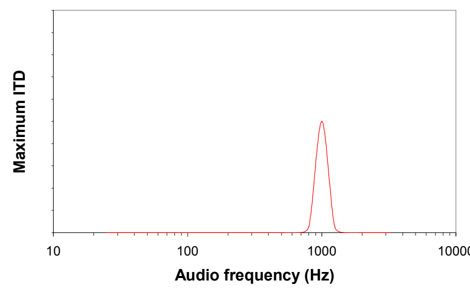


Figure 1b: Plot of audio frequency against maximum interaural time difference (ITD) of a narrow-band sample centred on 1000 Hz with a sinusoidal ITD fluctuation with a large maximum ITD

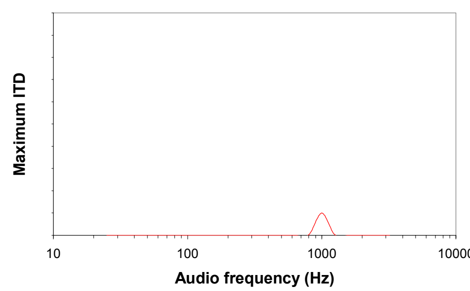


Figure 1c: Plot of audio frequency against maximum interaural time difference (ITD) of a narrow-band sample centred on 1000 Hz with a sinusoidal ITD fluctuation with a small maximum ITD

The first aim of the experiment can be expressed by which pair of audio samples in Figure 1 are perceived to have the most similar spatial properties. Two possibilities are considered.

Firstly, the subjective spatial effect may be independent of audio frequency. In that case the sample represented by Figure 1a would be perceptually most similar to the sample represented by Figure 1b.

Secondly, the subjective spatial effect may be dependent on audio frequency and a lower time magnitude of fluctuation is needed at higher audio frequencies to generate the most similar subjective spatial effect. In that case the sample represented by Figure 1a would be perceptually most similar to the sample represented by Figure 1c.

[1] For the purposes of clarity in this paper, all mention of the ITD fluctuation magnitude will be qualified with whether they are measured in terms of a peak interaural phase difference (phase magnitude) or a peak interaural time difference (time magnitude).

[2] The plots displayed here are representative for the purpose of explanation – the skirts of the plots of the narrow-band samples are representative of the frequency response rather than the maximum ITD values at frequencies either side of the centre frequency. The most important factor is the maximum ITD value at the centre frequency.

Aim 2: relationship between wide-band and narrow-band signals

The second aim of the experiment was to investigate the relationship between a wide-band sample with a given phase magnitude of sinusoidal ITD fluctuations and a narrow-band segment taken from the wide-band sample.

As discussed in [7], measurements of test signals replayed in real acoustic environments show that wide-band signals appear to have a greater magnitude of ITD fluctuations at lower audio frequencies. Because of this, for this experiment the wide-band samples were created with an ITD fluctuation magnitude of identical maximum phase difference at all audio frequencies. This creates a greater magnitude of maximum ITD values at lower audio frequencies compared to higher audio frequencies.

It was necessary to investigate what phase magnitude of sinusoidal ITD fluctuation of a wide-band sample was perceived to have the most similar spatial properties compared to a narrow-band sample with a given time magnitude of sinusoidal ITD fluctuations. A narrow-band sample may match to one of a number of wide-band samples with different phase magnitudes of ITD fluctuation as shown in Figure 2.

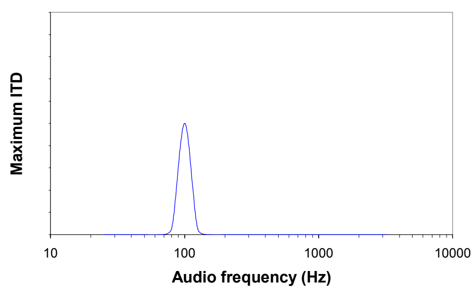


Figure 2a: Plot of audio frequency against maximum interaural time difference (ITD) of a narrow-band sample centred on 100 Hz with a sinusoidal ITD fluctuation with a large maximum ITD

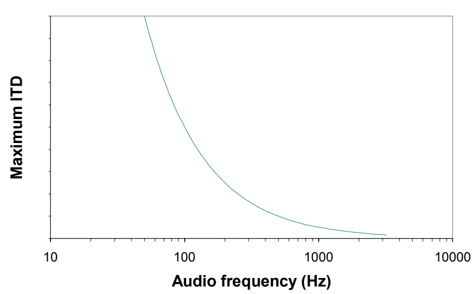


Figure 2b: Plot of audio frequency against maximum interaural time difference (ITD) of a wide-band sample with a sinusoidal ITD fluctuation with a large maximum ITD

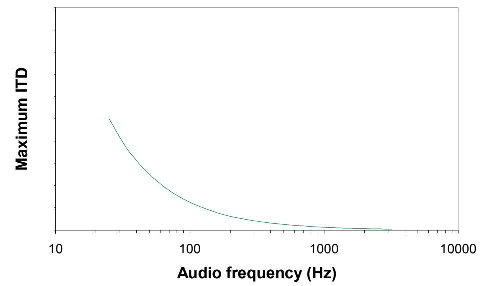


Figure 2c: Plot of audio frequency against maximum interaural time difference (ITD) of a wide-band sample with a sinusoidal ITD fluctuation with a medium maximum ITD

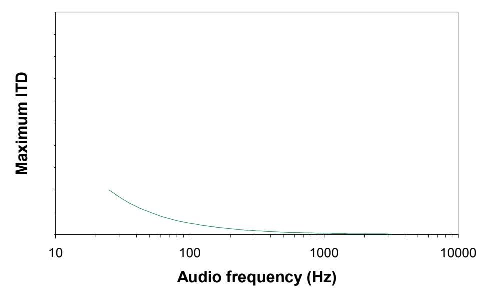


Figure 2d: Plot of audio frequency against maximum interaural time difference (ITD) of a wide-band sample Hz with a sinusoidal ITD fluctuation with a small maximum ITD

The second aim of the experiment can also be expressed by which pair of audio samples are perceived to have the most similar spatial properties, this time from Figure 2. Three possibilities are considered.

Firstly, the narrow-band sample (shown in Figure 2a) may be perceptually most similar to a wide-band sample of which the ITD fluctuation time magnitude matches at the frequency of the narrow-band sample (as shown in Figure 2b).

Secondly, the narrow-band sample (shown in Figure 2a) may be perceptually most similar to a wide-band sample of which the highest ITD fluctuation time magnitude matches the fluctuation time magnitude of the narrow-band sample (as shown in Figure 2c).

Thirdly, the narrow-band sample (shown in Figure 2a) may be perceptually most similar to a wide-band sample of which the integral of the ITD fluctuation time magnitudes across a given frequency range matches the integral of the ITD fluctuation time magnitudes of the narrow-band sample (as shown in Figure 2d).

It must be considered that the result may not match any of the possibilities discussed above. The possibilities outlined are for the purpose of explaining the aims of the experiment rather than attempting to predict the outcome of the experiment.

METHOD

The method chosen for this experiment was a ‘match to sample’ technique. The subjects were required to select one of a number of samples that a given stimulus was most similar to in terms of the specified subjective attributes[3]. This technique was chosen for a number of reasons.

Firstly, in subjective experiments there is always a potential problem of subjects misunderstanding linguistic descriptors used to describe perceptual attributes or scales [8]. This is especially problematic for the rather abstract noise samples used in this experiment. Also, as discovered in the previous experiment [7], individual subjects appear to perceive the ITD fluctuated noise samples slightly differently, which may result in a given linguistic descriptor being accurate for only some of the subjects. The match to sample technique means that there is less need for verbal attributes to be used by the experimenter. The samples themselves can be used to give details of the attributes under investigation, assuming that the samples only differ from each other in terms of that given attribute.

Secondly, when using scaling experiments there is the possibility that the subjects use the scales differently, or in a non-linear fashion [9]. The match to sample experiment is similar to a scaling experiment, with the samples making up the integer points of a scale. The samples are auditory markers along the scale, similar to the descriptive markers used in some experiments (such as ‘quite’ or ‘very’). The use of auditory examples as intermediate markers is arguably more accurate than using linguistic descriptors, as the audio samples are less open to interpretation compared to the rather vague terms used as linguistic markers. In addition, the audio samples form an absolute reference whereas the linguistic markers are often relative.

The use of the match to sample experiment did not entirely eliminate the need for linguistic descriptors. The subjects still needed instructions about which subjective attributes they should use as the basis for matching the stimulus to the samples. Whilst the differences between the samples were used as an indication of the relevant attributes, more precise instructions were needed to ensure that the subjects were judging the correct spatial features of the sound. The instructions were based on a mixture of the sketch-map drawing results from the previous experiment, and verbal descriptors interpreted from those drawings. The full instructions given to the subjects are shown in Appendix A.

The linguistic descriptors used in the instructions for the subjects were deliberately open to interpretation by the subject, depending on how they perceived the differences between the samples. For this reason a number of different pictoral descriptors were given. These were based on the results of the previous experiment. However, the key factors relating to the subjective attributes to communicate to the subjects were that they were to consider the spatial attributes of width and depth, and that they should compare the stimuli to the whole of the frequency range of the samples.

The subjects were also asked to listen to the samples given, and to determine the differences between them. They were encouraged to consider the differences and provide verbal descriptors of the differences they perceived. These verbal descriptors were recorded and reported as part of the previous experiment [7].

It was considered whether to allow the subjects to use a continuous grading scale, allowing the subject to judge that the stimulus was perceptually most similar to a non-existent sample that would lie between two of the samples given. However, the authors decided that this would dilute the simplistic nature of the match to sample experiment and make the task cognitively more complex for the subjects.

PROGRAMME MATERIAL

The wide-band samples were created in the same manner as those used in the previous experiment. They were 2-channel noise-like samples with a pre-determined sinusoidal ITD fluctuation similar to those used by Grantham and Wightman [3]. The samples were created using a large number of pairs of sine tones based on the equation below[4].

$$l = \sin[2\pi f_c t + \theta_c + m\sin(2\pi f_m t)]$$
$$r = \sin[2\pi f_c t + \theta_c - m\sin(2\pi f_m t)]$$

where l is the left ear signal
r is the right ear signal
f_c is the audio frequency
θ_c is a random phase component (identical in each ear)
m is the fluctuation phase magnitude
f_m is the fluctuation frequency

For every wide-band sample, the selection of values of f_c (audio frequency) were made in order to create a pink noise-like frequency response (equal power in each octave band). Values of f_c were used from 40 to 2560 Hz, a range of 6 octaves. The upper limit was based on the previous research that suggests that audio frequencies above a few kilohertz do not alter the magnitude of the spatial effect [10]. A large number of pairs of sine tones with random starting phases (θ_c) were used in order to minimise the periodic interaction of the sine tones with each other that cause monaural amplitude fluctuations. This was a trade-off between minimising the amplitude fluctuations and the time required to create the samples. Therefore there were 1000 pairs of frequency modulated sine tones in each octave.

The experiment used a fixed value of f_m of 100 Hz, as the purpose of the experiment was to evaluate the subjective effect of ITD fluctuations at fluctuation frequencies above the rate at which the effect of binaural sluggishness causes the perception of a stationary spatial impression. In the previous experiment [7] the samples with a fluctuation frequency of 100 Hz were judged to create a subjective spatial effect which contained no moving components.

If more than one value of f_m had been used, the subject would have had to familiarise themselves with two or more sets of samples, therefore greatly increasing the complexity of the task. The use of a single value of f_m may have limited the validity of the experiment at other fluctuation frequencies, however this was required to enable the production of more accurate results.

The values of m were chosen to cover a range of the subjective spatial effect. The lowest value was 0.0, giving a baseline value of no ITD fluctuations. The perceived spatial dimensions of this wide-band stimulus were identical to a mono pink noise stimulus fed to both ears. The highest value of m that was practical was 1.5. If values much larger than this were used then unwanted side-effects were created which significantly altered the timbre of the samples.

The maximum ITD or time magnitude of the samples by audio frequency can be expressed by the following equation.

$$\text{Maximum ITD} = 2m/2\pi f_c \quad [7]$$

[3] For the purposes of clarity in this experiment, the narrow-band audio excerpts that were presented singly are termed the stimuli, and the wide-band audio excerpts that these stimuli were matched to are termed the samples.

[4] For a more detailed discussion of the method of creating the stimuli noise samples with sinusoidal ITD fluctuations, the reader is referred to [3, 7]

From this the maximum ITD or time magnitude of the sample with the largest fluctuation phase magnitude (a value of m = 1.5) can be calculated for the audio frequency range it spans. This is shown in Figure 3 below.

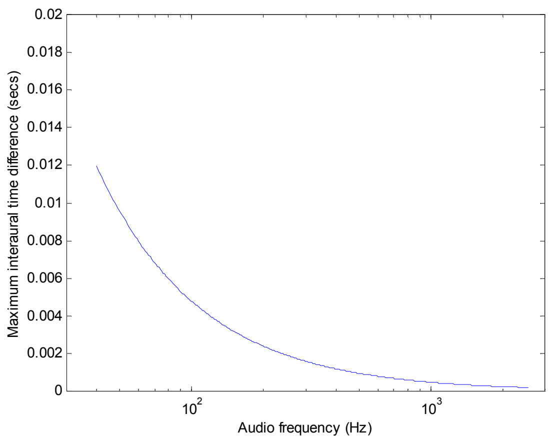


Figure 3: Plot of the maximum ITD against audio frequency for the sample with the largest ITD fluctuation phase magnitude (m = 1.5).

The intermediate values between the sample with the lowest ITD fluctuation phase magnitude (m = 0.0) and the sample with the highest ITD fluctuation phase magnitude (m = 1.5) had to be chosen with care. There was a trade-off between two main factors. Firstly, not having enough samples in the range would cause a quantisation of the results due to insufficient resolution in the step size, meaning that there was a large perceptual distance between each of the samples. Secondly, if there were too many samples in the range then the task would be more difficult for the subjects. In addition, there could be too small a perceptual difference between a number of samples (in other words a number of samples lying within the just noticeable difference range of each other), therefore adding to the noise of the experiment results.

The spacing of the ITD fluctuation phase magnitudes of the samples also had to be decided. As there was no data available regarding the linearity of the subjective spatial effect with increasing ITD fluctuation phase magnitude, a linear perceptual scale could not be determined. Therefore the spacing of the ITD fluctuation phase magnitudes was based on a linear increase in values of m.

After critical listening by the authors, it was decided to use 16 samples, from a value of m = 0.0, in steps of 0.1 to a maximum value of m = 1.5. This appeared to be an acceptable compromise between the factors mentioned above.

The narrow-band stimuli were created by filtering the wide-band samples using a 1/3rd-octave band-pass filter. The centre frequencies were chosen to be at 100 Hz and 1000 Hz, giving a practical range of frequencies within the bandwidth of the wide-band samples. In addition, a third centre frequency of 500 Hz was chosen. The results from this frequency should give some indication of whether the trend between the two outside frequencies is linear.

The ITD fluctuation magnitudes of the narrow-band stimuli were chosen to enable accurate testing of the two most likely outcomes of the first aim of the experiment as discussed in the previous section.

For the first possibility, there had to be the option that at least one of the narrow-band stimuli of each audio frequency would have a very similar maximum ITD value or time magnitude to at least one of the narrow-band stimuli of the other audio frequencies. If these stimuli with similar maximum ITD values were judged to be subjectively similar, then it would indicate that the first possibility was correct. For the second possibility, there had to be the option that at least one of the narrow-band stimuli of each audio frequency range would have a lower maximum ITD than at least one of the narrow-band stimuli in a lower audio frequency range. The ratio between the maximum ITD values in each frequency range would ideally be related by phase. If these stimuli with an identical maximum phase difference were judged to be subjectively similar, then it would indicate that the second possibility was correct.

The authors were careful not to exclude the possibility of other conclusions in the experiment. Any other outcomes were not specifically catered for within the design, however if none of the sets of stimuli were judged to be subjectively equal (either by time magnitude / maximum ITD or phase magnitude / values of m), then it would indicate that the two outlined possible outcomes were incorrect. This would require further investigation in an additional experiment.

In order to examine the second aim of the experiment as discussed in the previous section, the narrow-band stimuli were created by filtering certain wide-band samples that were used in the experiment. This enabled an examination of whether the subjects judged the stimuli to be subjectively equal to the wide-band samples they were filtered from.

It was decided to use three values of m to give three levels of ITD fluctuation phase magnitude. Each of these were filtered at the three different audio frequencies to give nine stimuli, each with a different maximum ITD of sinusoidal fluctuation. Using the same value of m for the three different audio frequencies meant that for each level of fluctuation phase magnitude, the three stimuli at the three different audio frequencies were related by equal maximum phase deviation. The three values of m were chosen to give a wide range of values, and give approximately the same maximum ITD value or fluctuation time magnitude for one in each of the frequency ranges. This meant that the fluctuation time magnitude of the lowest audio frequency stimulus with the lowest ITD fluctuation phase magnitude was approximately equal to the fluctuation time magnitude of the middle audio frequency stimulus with the medium ITD fluctuation phase magnitude. This was again approximately equal in fluctuation time magnitude to the highest audio frequency stimulus with the highest ITD fluctuation phase magnitude. This is shown in Table 1.

Fluctuation phase magnitude	**Maximum fluctuation time magnitude (msec)**		
	100 Hz	500 Hz	1000 Hz
0.1	0.32	0.064	0.032
0.7	2.2	0.45	0.22
1.3	4.1	0.83	0.41

Table 1: Table of the maximum ITDs or fluctuation time magnitude (in msec) of the nine stimuli used in the experiment, organised by the centre frequency of the 1/3rd octave filter, and the fluctuation phase magnitude (m) of the wide-band sample from which it is filtered.

By using such a large range of values of m for the narrow-band stimuli compared to the range of m of the wide-band samples, it was expected that some results may be affected by ceiling or floor effects. However, this was not expected to be a significant problem as there should be sufficient unaffected data points to discover any trends in the data.

To summarise, there were 9 stimuli to be auditioned and matched to one of 16 samples. The nine stimuli were made up of two independent variables, audio frequency and fluctuation frequency.

EXPERIMENTAL SET-UP

Physical set-up

The experiment was run on a PC containing a soundcard with an S/PDIF output. The S/PDIF output was plugged into a Yamaha 02R mixer for conversion to AES/EBU, and then connected to a Tascam DA-30 for D/A conversion and headphone amplification. The signals were replayed over Sennheiser HD-545 headphones.

The audio extracts were replayed at 90dB SPL linear average as measured with the headphones positioned on a KEMAR head with occluded ear simulators. The replay level was chosen based on listening. The loudness was increased until the samples could be heard clearly by the authors without having to strain to hear the detailed effects. This was also consistent with the replay level used in the previous experiment.

User interface

The user interface needed to display audio playout controls for the stimulus and the 16 samples, allow simple switching between them, and enable many randomised stimulus orders to be easily created.

A number of different methods for creating the user interface were evaluated and HTML was chosen as the easiest solution. The user interface was written in HTML and Javascript, using embedded Microsoft Windows Media Player components to play the audio extracts. For each presentation the stimulus playout controls were located in the upper section of the page, whilst the samples were laid out in the lower section of the page. The samples were labelled A to P. This alphabetical labelling was used so that the subjects were not subconsciously led into the assumption that the perceptual scale of the samples was necessarily linear, as might have been assumed from numerical labelling.

The samples were always laid out in the same order for the subjects, in incremental steps from the lowest phase magnitude of ITD fluctuation at A, to the highest phase magnitude of ITD fluctuation at P. This may have led to the problem that the subjects did not listen to the samples as carefully towards the end of the experiment as they became more familiar with the samples. However, the alternative of randomising the sample orders would have made the task a lot more difficult for the subjects. Firstly, they would have had to listen to all the samples for each separate stimulus presentation. Secondly, they would have had to remember where each sample was for the purposes of comparison. The authors felt that the cognitive load imposed by this randomisation would have increased the chances of error and the amount of listener fatigue.

The interface was run using Microsoft Internet Explorer. The user interface is shown in Figure 4.

The Media player components were configured so that clicking on play would start the associated audio excerpt. Clicking on the play icon of another audio excerpt would silence the audio excerpt already playing and start the new audio excerpt. The subjects could listen to the audio extracts in any order, and could repeat them as many times as they required.

The judgement was recorded on paper, using an answer sheet with a similar layout to the user interface. An example can be seen as part of the instructions given to subjects in Appendix A. The stimulus number on the user interface was the chronological order number that corresponded to the number on the answer sheet.

Loudness alignment

The loudness of an auditory signal can have a large effect on how it is perceived [11]. If the loudness of the samples and stimuli were not aligned, it would be a confounding variable in the experiment. This means that any perceived differences between two stimuli may partially be due to the loudness differences, in addition to the controlled changes to the independent variables. As the experiment involved comparing all the stimuli with the same set of samples, it

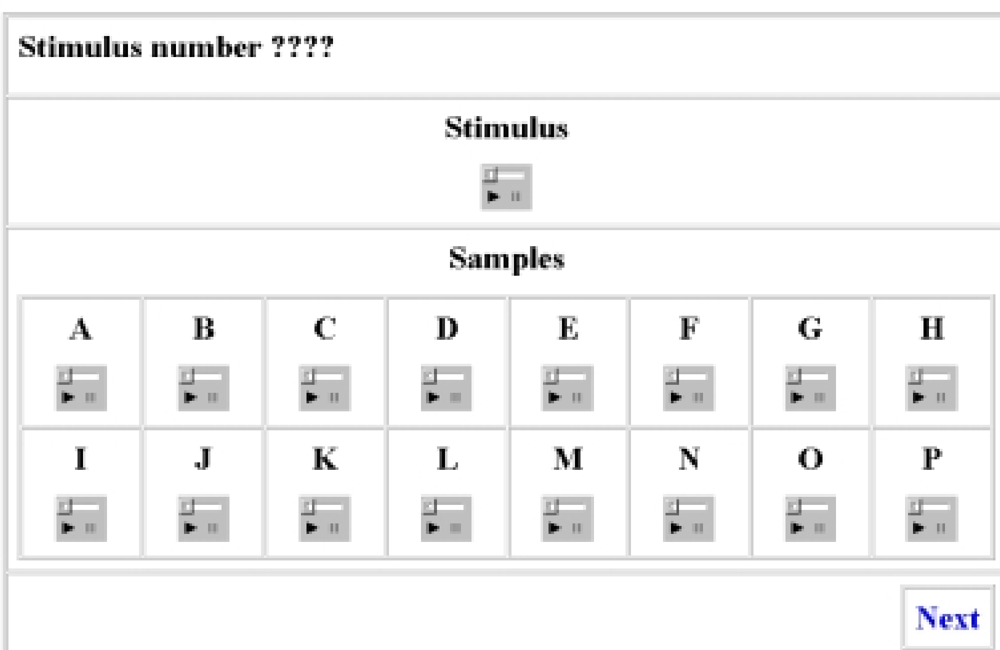


Figure 4: The user interface created for the match to sample experiment. The stimulus is given at the top of the screen and the 16 samples are in the lower half of the screen. The subjects were asked to select the single sample that they considered to be most similar to the stimulus in terms of the given subjective attributes.

	100Hz	500 Hz	1000 Hz
Moore model	+21.5	+10.3	+8.5
Average power	-2.5	-0.8	-0.3
Dolby loudness	+6.0	+4.0	-1.0
Mean of subjective judgements (outliers removed)	+2.3	+1.5	+0.2
Max of subjective judgements (outliers removed)	+7.2	+6.8	+1.2
Min of subjective judgements (outliers removed)	-2.8	-2.8	-2.0

Table 2: Approximate gain required to make the narrow-band stimuli the same loudness as the wide-band samples as determined by the different alignment or measurement techniques

was necessary to align all the samples and stimuli to have the same subjective loudness.

All the samples and stimuli were measured using the Moore loudness model [12]. The results from this showed that within groups of similar audio extracts, the samples and stimuli were very similar in loudness. This meant that within each group no loudness alignment was needed. There were four such groups: the first containing all the wide-band samples; the second containing all the narrow-band stimuli centred on an audio frequency of 100 Hz; the third containing all the narrow-band stimuli centred on an audio frequency of 500 Hz; and the fourth containing all the narrow-band stimuli centred on an audio frequency of 1000 Hz.

However, between each of the groups there was a large difference in measured loudness. The audio extracts in each of the groups were aligned to measure the same loudness according to the Moore model. The results of the loudness measurement are shown in Table 2.

When the audio extracts aligned by this method were auditioned, it was apparent that the narrow-band stimuli sounded far too loud when compared to the wide-band samples. The difference in loudness between the narrow-band stimuli centred on an audio frequency of 100 Hz and the wide-band samples were most extreme, though the difference between the other stimuli and the wide-band samples were significant.

With the failure of this loudness alignment technique, other level alignment techniques[5] were tried, including equalising for average power, and using a Leq(m) measurement using a Dolby level meter. The results of the different alignment techniques are shown in Table 2. As all three objective alignment measurement techniques gave very different results, it was uncertain which, if any, could be trusted. For this reason a small subjective experiment was carried out where subjects adjusted the level of one stimulus from each of the narrow-band stimulus groups so they were the same loudness as one of the samples from the wide-band sample group.

The audio extracts were subjectively loudness aligned in pairs. Each extract was looped, and the subjects could switch between them at any time, taking as long as they needed. The audio extracts chosen for comparison were the samples from each group with the minimum magnitude of ITD fluctuations. These were chosen to eliminate the possibility that large ITD fluctuations affect the perception of loudness.

Ten subjects took part in the test. The results showed a large difference between the subjects, a range of –2.8 to +7.2 dB once the outliers were removed by the use of a boxplot. Even though the range was large, it was considerably less than the difference indicated by the Moore loudness model. The results of the different alignment techniques are shown in Table 2.

The large difference between the results of the individual subjects could be due to a number of factors. Firstly, because of the large difference in the frequency ranges of the different audio extracts, it may be that the subjects found it difficult to make the loudness of the wide-band and narrow-band extracts equal, therefore causing a large amount of error. Secondly, it may be that the subjects were comparing different frequency ranges of the wide-band samples when aligning the loudness. Depending on the strategy used by the subject they may have compared the whole of the sounds, or specific frequency ranges, therefore giving different results.

Thirdly, it may be that there were changes in the loudness of the audio extracts over the duration of the extracts. Depending on which part of the extract the switch between the pair of audio extracts was made, the perceived loudness may have been different. Finally, the difference may have been due to each individual having a different ear frequency response. It may be that each subject has a different sensitivity to sounds at different frequencies, therefore needing a different amount of level of a narrow-band sound at certain frequencies to make them sound equally loud compared to a wide-band sound.

If it is the case that the subjects judge the narrow-band stimuli to have very different subjective loudness, then it may be that the subjects should be allowed to make their own loudness alignment judgements prior to the main experiment. This was considered, but rejected because it would introduce an additional confounding variable into the experiment design. In addition, it would make the test more complicated for the subject.

There still was a need to make a single alignment for all subjects to use. As none of the objective alignment techniques appeared to be most appropriate in a way that could be justified, the means from the subjective experiment with the outliers removed were used to align the four groups of samples and stimuli.

EXPERIMENT PROCEDURE

The subjects were given the instructions shown in Appendix A. As mentioned above, the subjects were then required to listen to the samples, and to consider the differences between them. To encourage this, the subjects were asked to describe the differences between the samples in their own words.

There were 9 stimuli to judge in the experiment. In order to evaluate the consistency of the subjects, all the stimuli were auditioned twice. The stimuli were presented in two sets of nine, with a different random order each time. In an attempt to cancel out any perceptual effects due to specific orders of stimuli presentation, each subject also had a different random order of the stimuli. The stimuli were presented individually, together with the 16 samples. The subject was required to select the sample that matched the stimulus most closely in terms of the given spatial attributes, then move onto the next stimulus.

The listening tests were carried out in a darkened ITU-R BS.1116 [13] standard listening room in order to limit extraneous distractions, both audible and visual.

[5] It must be noted that whilst the Moore model and the subjective judgements are loudness measures, the equal power and Dolby Leq(m) measurements are more simple level measurements.

22 subjects undertook the listening tests. They were all final year students, postgraduate students or staff at the Institute of Sound Recording at the University of Surrey. Therefore they can be considered to be expert listeners. The subjects were not informed in any way about what they would be listening to, apart from the facts contained in the instructions in Appendix A. The duration of each listening was on average 20 minutes.

RESULTS AND DISCUSSION

The results of the experiment were in terms of the sample that each stimulus was matched to. This was an alphabetical value of between A and P. For the purposes of mathematical analysis this was converted into a numerical value from 1 to 16, representing the alphabetical labels in numerical order. For the rest of the analysis this will be referred to as the 'subjective judgement'.

There were two independent variables that could be entered into the test. Firstly, the audio frequency of the stimulus (one of three values: centred on 100 Hz; centred on 500 Hz; or centred on 1000 Hz). Secondly, the fluctuation phase magnitude of the stimulus (one of three values of m: 0.1; 0.7; or 1.3). Finally, the individual subject could be entered into the analysis.

The first stage of the analysis was to check that the data conformed to the assumptions of ANOVA. This analysis showed that due to floor effects, the data were not normally distributed and did not have homogeneous variance. In addition, the use of the match to sample technique meant that the data contained only integer points, and it could not be assumed that the integer points were a linear scale.

Because of this, the validity of applying the ANOVA model was tested by comparing the results to analysis carried out using a non-parametric Kruskal-Wallis test. The significance value results of a univariate ANOVA and a one-way non-parametric Kruskal-Wallis test can be compared in Table 3.

Independent variable	ANOVA	Kruskal-Wallis
Subject	0.157	0.153
Audio frequency	0.337	0.377
Fluctuation phase magnitude	0.000	0.000

Table 3: Table of the significance value results of both univariate ANOVA and one-way Kruskal-Wallis tests of the data from all the subjects using the independent variables of Subject, Audio frequency and Fluctuation phase magnitude.

It is apparent that the results from the ANOVA and the Kruskal-Wallis tests were very similar. Therefore, it could be assumed that the ANOVA was giving reasonable results for this data set and could be used to analyse the data further.

The results from each subject were examined, and the consistency of each subject was measured by calculating the mean absolute difference between the two presentations of each stimulus. This showed that a number of subjects were very inconsistent with their judgements. Therefore it was decided to filter out the most inconsistent subjects as their widely differing results were unreliable and could significantly alter the averages of any analysis.

The data was analysed with the results of each subject being removed individually in the order of consistency, from worst consistency to best. This showed that there was no statistically meaningful value beyond which to discard results based on a large change in the significance or error values from the ANOVA. Therefore the cut-off point was arbitrarily chosen to select only the subjects with a mean absolute difference between the two presentations of each stimulus of less than 2.5. This seemed to the authors to be a reasonable value to choose. Using this value to select the most consistent subjects resulted in there being 11 sets of results for further analysis.

The data was then re-analysed by univariate ANOVA and one-way non-parametric Kruskal-Wallis tests to retest the validity of the ANOVA model. The results, shown in Table 4, show that again the results of the ANOVA and the Kruskal-Wallis tests were very similar, meaning that the ANOVA could be used to analyse the data further.

Independent variable	ANOVA	Kruskal-Wallis
Subject	0.521	0.494
Audio frequency	0.627	0.598
Fluctuation phase magnitude	0.000	0.000

Table 4: Table of the significance value results of both univariate ANOVA and one-way Kruskal-Wallis tests of the data from the 11 selected subjects using the independent variables of Subject, Audio frequency and Fluctuation phase magnitude.

The data from the 11 selected subjects were then entered into a type III sum of squares general linear model ANOVA with the fixed factors of Subject (SUBJECT), Audio frequency (FREQ) and Fluctuation phase magnitude (FLUCTMAG) and all interactions. The results are shown in Table 5.

All of the factors and interactions were significant beyond the $p = 0.05$ level. However, it is apparent that the partial eta squared value (an indication of the effect size of each factor and interaction [14]) was largest for the Fluctuation phase magnitude. This was followed by the Subject and Audio frequency interaction and then the Subject factor.

It was therefore the case that the change of Fluctuation phase magnitude was causing the largest perceived difference to the subjective results. There was a significant difference between the results of the individual subjects, as may be expected in a subjective experiment. The interactions show that the subjects were judging the effects of the Audio frequency and the Fluctuation phase magnitude slightly differently. However, compared to the large difference caused by the Fluctuation phase magnitude this was a minor difference and could therefore be reasonably disregarded.

The significance of the Audio frequency factor could also reasonably be disregarded, as the interaction of Audio Frequency and Subject had a larger significance. This meant that the differences in the results between the different Audio Frequencies were predominantly subject-dependent, and therefore not the same for the whole experiment population.

In order to view the main results of the aims of the experiment as described in the Aims section above, the means and associated 95% confidence intervals of the subjective judgements needed to be viewed separated by stimulus. Figure 5 shows the subjective judgements of the nine stimuli plotted against the fluctuation time magnitude of the stimuli. The stimuli are labelled by the centre frequency of the filter used to create them from the wide-band samples. The fluctuation time magnitude axis is plotted on a logarithmic scale for easier viewing.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4721.040	98	48.174	18.134	.000	0.947
Intercept	14017.960	1	14017.960	5276.722	.000	0.982
SUBJECT	232.485	10	23.248	8.751	.000	0.469
FREQ	23.798	2	11.899	4.479	.014	0.083
FLUCTMAG	3581.253	2	1790.626	674.038	.000	0.932
SUBJECT * FREQ	417.424	20	20.871	7.856	.000	0.613
SUBJECT * FLUCTMAG	205.303	20	10.265	3.864	.000	0.438
FREQ * FLUCTMAG	37.172	4	9.293	3.498	.010	0.124
SUBJECT * FREQ * FLUCTMAG	223.606	40	5.590	2.104	.002	0.460
Error	263.000	99	2.657			
Total	19002.000	198				
Corrected Total	4984.040	197				

R Squared = .947 (Adjusted R Squared = .895)

Table 5: Table of results from a type III sum of squares general linear model ANOVA using the subjective judgement data from the 11 selected subjects with the independent variables of Subject (SUBJECT), Audio frequency (FREQ) and Fluctuation phase magnitude (FLUCTMAG) entered as fixed factors with all interactions.

The results are given with respect to the aims set out above. The first aim was to discover whether the magnitude of the subjective spatial effect of an audio signal with specific ITD fluctuations was dependent on the audio frequency.

The subjective judgement was a scale related to the fluctuation magnitude of the samples to which each stimulus was matched, and the stimuli varied only in Fluctuation magnitude and Audio frequency. Therefore it would be logical that if the magnitude of the subjective effect was independent of audio frequency and the perceived effect was similar for the wide-band samples and narrow-band stimuli, then there would be a monotonically increasing trend through all the stimuli. The results in Figure 5 show this to be not the case.

The experiment was designed such that if the subjective effect was

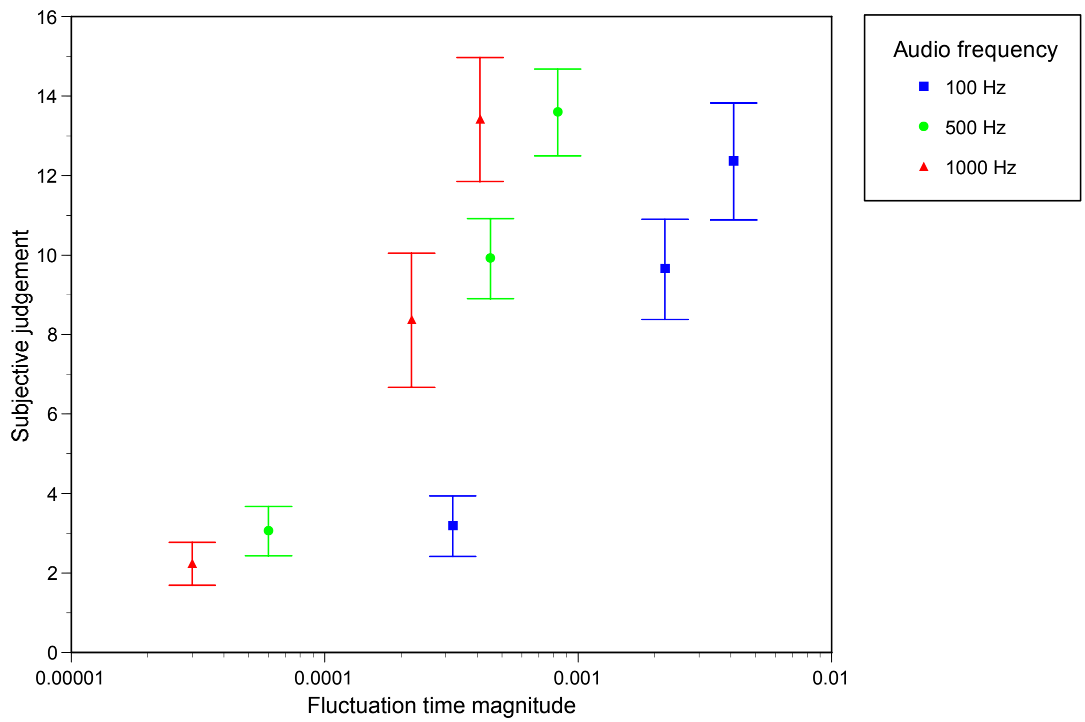


Figure 5: Plot of the means and associated 95% confidence intervals of the subjective judgement data from the 11 selected subjects of the nine stimuli against the fluctuation time magnitude, labelled by the audio frequency of the stimuli.

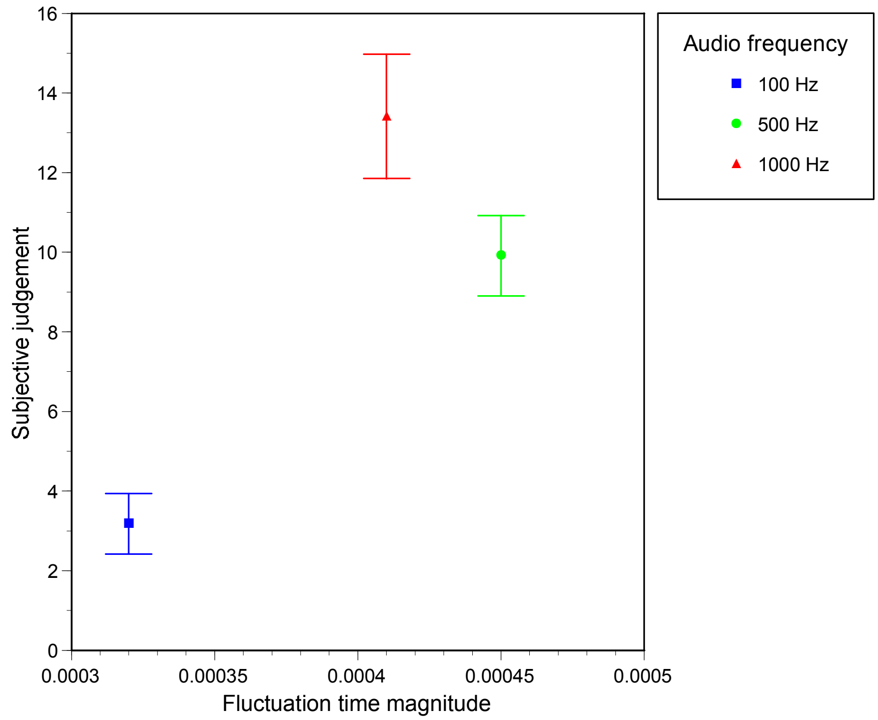


Figure 6: Plot of the selected results of the stimuli with approximately equal fluctuation time magnitude showing the means and associated 95% confidence intervals of the subjective judgement data from the 11 selected subjects plotted against the fluctuation time magnitude, labelled by the audio frequency of each stimulus.

not dependent on the audio frequency, then the lowest audio frequency stimulus with the lowest ITD fluctuation phase magnitude would be approximately equal to the middle audio frequency stimulus with the medium ITD fluctuation phase magnitude. This in turn would be approximately equal to the highest audio frequency stimulus with the highest ITD fluctuation phase magnitude. These results are the centre three error bars in Figure 5, and are shown in Figure 6.

It is clear from the results in Figure 6 that these three stimuli were judged to be significantly different from each other. Therefore it can be concluded that in this case, the magnitude of the subjective spatial effect created by ITD fluctuated audio signals was dependent on the audio frequency of the signal.

Having found that the magnitude of the spatial effect was dependent on the audio frequency, the data was investigated to uncover whether there was a consistent relationship between the audio frequency and the magnitude of the spatial effect. It is apparent from Figure 5 that there was an approximately linear trend with increasing fluctuation time magnitude within each group of stimuli with the same audio frequency. In addition, across all the stimuli there were groups of stimuli with similar results that appeared to match the fluctuation phase magnitude of the stimuli.

This can be seen more clearly if the means and associated 95% confidence intervals are plotted for the stimuli by Fluctuation phase magnitude and Audio frequency. This is shown in Figure 7.

It is apparent from Figure 7 that apart from one overlapping error bar, each group of stimuli with the same fluctuation phase magnitude was significantly different from the groups of stimuli with different fluctuation phase magnitudes. In addition, within each group of stimuli with the same fluctuation phase magnitude, the stimuli were not significantly different by audio frequency. Therefore, it appears that the perceived magnitude of subjective effect was approximately the same for a given maximum phase difference of ITD fluctuations.

The second aim of the experiment was to examine the relationship between a wide-band sample with a given phase magnitude of sinusoidal ITD fluctuations and a narrow-band segment taken from the wide-band sample. The values of m used for the stimuli were 0.1, 0.7 and 1.3. These equate to the wide-band samples in the experiment labelled as B, H and N, and these in turn equate to the integer points of 2, 8 and 14 in the numerical subjective judgement results.

The relationship was then examined by comparing the subjective judgements of the stimuli at each level of fluctuation phase magnitude to the value of the subjective judgement that equates to the sample from which the stimulus was filtered. This is shown in Figure 8.

The confidence intervals for the stimuli with the lowest ITD fluctuation phase magnitude ranged from approximately 1.5 to approximately 4. These were all filtered from the sample that was at the integer scale point equating to 2. The confidence intervals for the stimuli with the medium ITD fluctuation phase magnitude ranged from approximately 6.5 to approximately 11. These were all filtered from the sample that was at the integer scale point equating to 8. The confidence intervals for the stimuli with the highest ITD fluctuation phase magnitude ranged from approximately 11 to approximately 15. These were all filtered from the sample that was at the integer scale point equating to 14.

It can be seen that for the two lower fluctuation phase magnitudes there was a trend for the narrow-band stimuli to be perceived to have a larger spatial effect than the sample from which they were filtered, especially for the lower two audio frequencies. However, this was not the case for the highest fluctuation phase magnitude, where the trend was for the narrow-band stimuli to have a smaller perceived

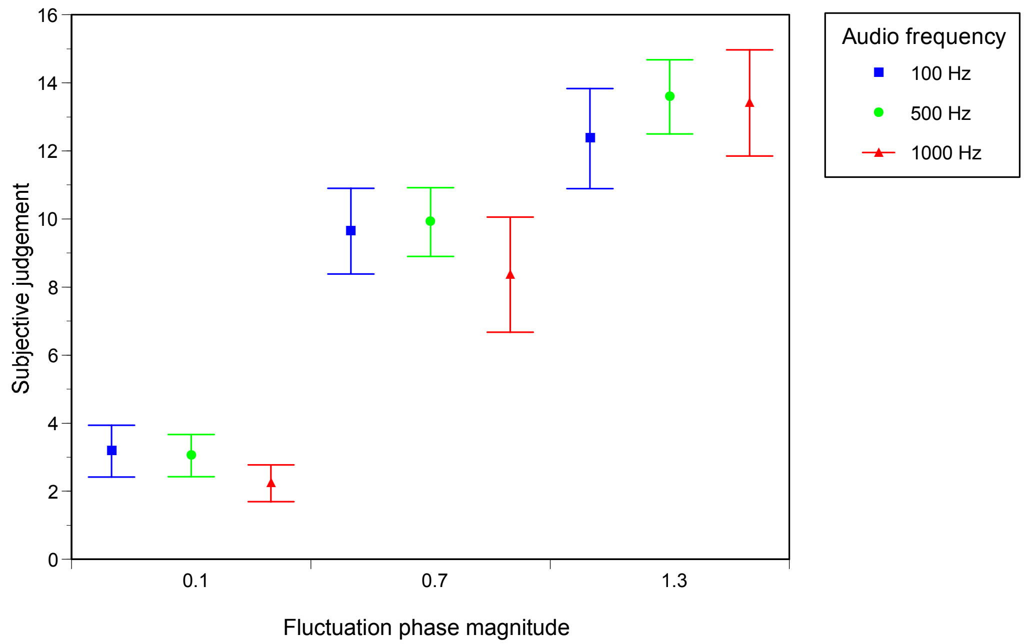


Figure 7: Plot of the means and associated 95% confidence intervals of the subjective judgement data from the 11 selected subjects, separated by Audio frequency and Fluctuation phase magnitude.

spatial effect than the sample from which they were filtered, especially for the lowest audio frequency.

The trend that the narrow-band stimulus was judged to have a larger spatial effect than the sample from which it was filtered may be due to the loudness of the audio extracts. For this experiment, the stimuli were all loudness aligned with the samples by use of subjective loudness alignment. This meant that the narrow-band stimuli with the lowest audio frequency were amplified most, followed by the stimuli with the middle audio frequency, with the stimuli with the highest audio frequency hardly changed. This trend appears to be the inverse of the results shown in Figure 8.

Therefore it may be that increasing the amplitude of the narrow-band samples has caused this effect, though it was needed to give equal subjective loudness for all the audio extracts.

The opposite trend can be seen in the results of the stimuli with the

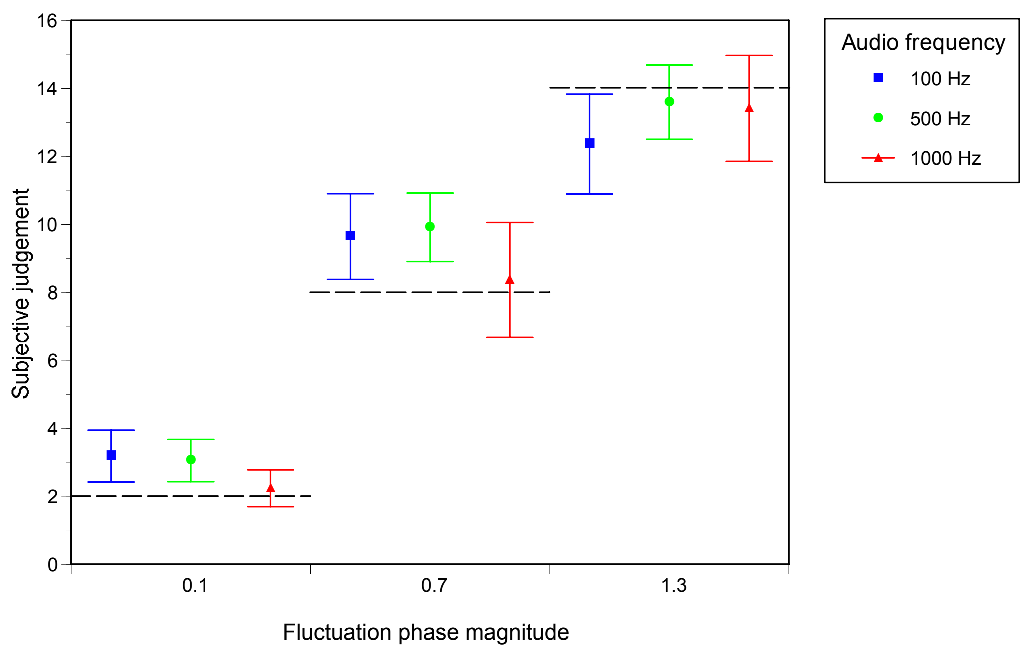


Figure 8: Plot of the means and associated 95% confidence intervals of the subjective judgement data from the 11 selected subjects, separated by Audio frequency and Fluctuation phase magnitude, with the addition of dotted lines showing the subjective value equivalent to the sample from which each group of stimuli was filtered.

largest fluctuation phase magnitude. This may have been due to the subjective effect starting to compress at large ITD fluctuation time magnitudes. This would have caused the lower audio frequencies to have a lesser subjective effect compared to the higher audio frequencies, as the lower audio frequencies have a larger time magnitude of ITD fluctuations, and therefore the subjective effect for these frequencies would be compressed more than the higher frequencies with smaller maximum ITDs.

CONCLUSIONS

The first question set out in the Aims section above was to find whether the magnitude of the subjective spatial effect created by audio signals with specific interaural time difference (ITD) fluctuations is dependent on the audio frequency. In this experiment it was found that:

- the magnitude of the subjective spatial effect created by audio signals with specific ITD fluctuations was dependent on the audio frequency
- the relationship between the audio frequency and the magnitude of ITD fluctuations to create a uniform magnitude of subjective spatial effect appears to be based on the maximum phase difference of the ITD fluctuations

The second question set out in the Aims section above was to investigate the relationship between a wide-band sample with a given phase magnitude of sinusoidal ITD fluctuations and a narrow-band segment taken from the wide-band sample. In this experiment it was found that:

- for the specific audio signals used in this experiment with a lower maximum ITD, the relationship between a wide-band sample with a given phase magnitude of sinusoidal ITD fluctuations and a narrow-band segment taken from the wide-band sample was that the narrow-band stimuli were perceived as having a slightly larger subjective spatial effect, possibly due to the loudness difference at different audio frequencies
- for the specific audio signals used in this experiment with a higher maximum ITD, the relationship between a wide-band sample with a given phase magnitude of sinusoidal ITD fluctuations and a narrow-band segment taken from the wide-band sample was that the narrow-band stimuli were perceived as having a slightly smaller subjective spatial effect, possibly because of the compression of the subjective effect at higher maximum ITDs

The results of this experiment can now be applied back into a measurement based on calculating the magnitude of the ITD fluctuations. Firstly, it appears that the magnitude of the ITD fluctuations should either be measured as a maximum phase difference as opposed to a maximum ITD, or the maximum ITD should be weighted by audio frequency to give the same result. Secondly, the measurement should be calculated so that a narrow-band stimulus causes a similar result to a wide-band stimulus if they have ITD fluctuations with the same frequency and the same maximum phase difference, though the loudness and the compression effects at high time magnitudes may have to be taken into account.

FURTHER WORK

There are a large number of parameters that need to be investigated to fully understand the subjective effect of ITD fluctuations and the refinement of measurements based on them. However, specific limitations related to this experiment are the following.

Firstly, it is uncertain whether the effect is the same at all ITD fluctuation frequencies. Whilst the authors hypothesise that this is indeed the case for the range of ITD fluctuation frequencies which cause a similar spatial effect (i.e. the perception of a stationary wide diffuse source with no moving components), this should be investigated further.

Secondly, it is uncertain whether the subjective effect is the same for more complex ITD fluctuations than single sinusoidal fluctuations. It is assumed that the subjective effect would be similar, though this is not necessarily the case and needs to be investigated.

Thirdly, the results appear to indicate that the magnitude of the subjective effect of the ITD fluctuations is somewhat dependent on the loudness of the audio signal. This need to be investigated.

Finally, it is uncertain whether the subjective effect is the same for ITD fluctuations delivered over loudspeakers with a perceivable source in front of the subject. This will be investigated in following experiments.

ACKNOWLEDGEMENTS

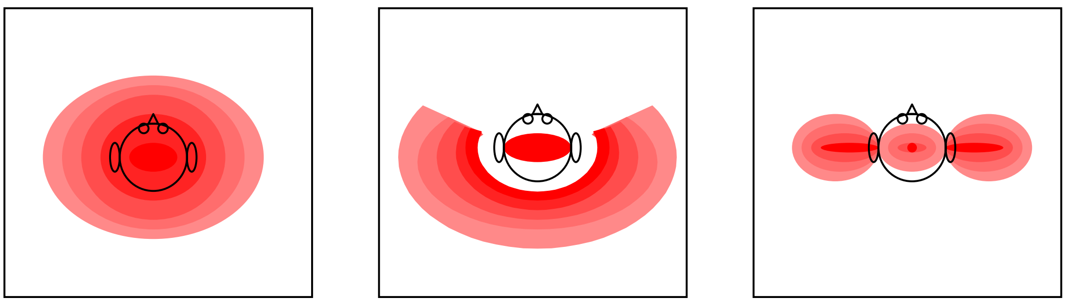
The authors would like to thank David Meares and BBC R&D for sponsorship and provision of equipment, the MEDUSA team for their thought-provoking discussions and guidance, the AES Educational Foundation for financial assistance and Lin Oskam for her patient proof reading.

APPENDIX A: INSTRUCTIONS FOR LISTENERS

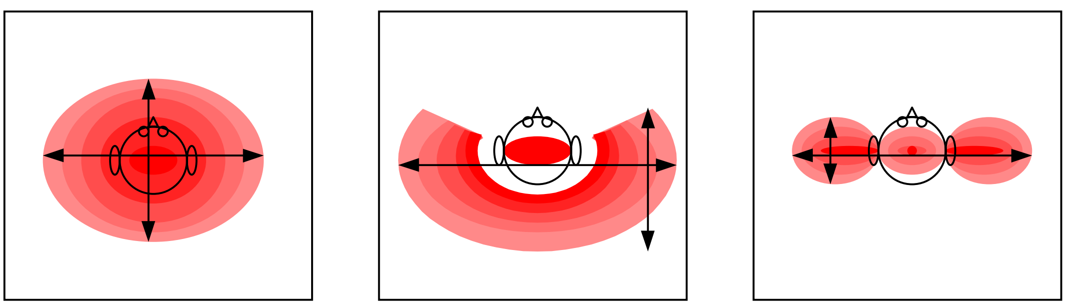
The purpose of this listening test is to judge the subjective spatial dimensions of a number of audio extracts. The listening test is a ‘match to sample’ experiment. You are asked to select one of a number of samples that matches the stimulus most closely in terms of its spatial attributes. There are sixteen samples of which you must choose one. The spatial attributes of each of the samples are slightly different and vary incrementally from the first (A) to the last (P). The same 16 samples are used for all the stimuli judgements, and they are always arranged in the same order. At the beginning of the test is an opportunity to familiarise yourself with the samples. Please take the time to listen to all of them closely and to try and determine any spatial differences you can perceive between them.

The stimuli are narrow-band and of various frequencies whilst the samples that you are matching them to are wide-band noise samples. You are asked to match the subjective dimensions of the whole of the sample bandwidth, not just the same frequency range as the stimulus.

The subjective spatial dimensions that are most important for this experiment are the maximum size of the sound, the maximum width and depth that you perceive the sound to cover. For example, you may perceive the sounds to have spatial positions and dimensions similar to the one of the following diagrams.



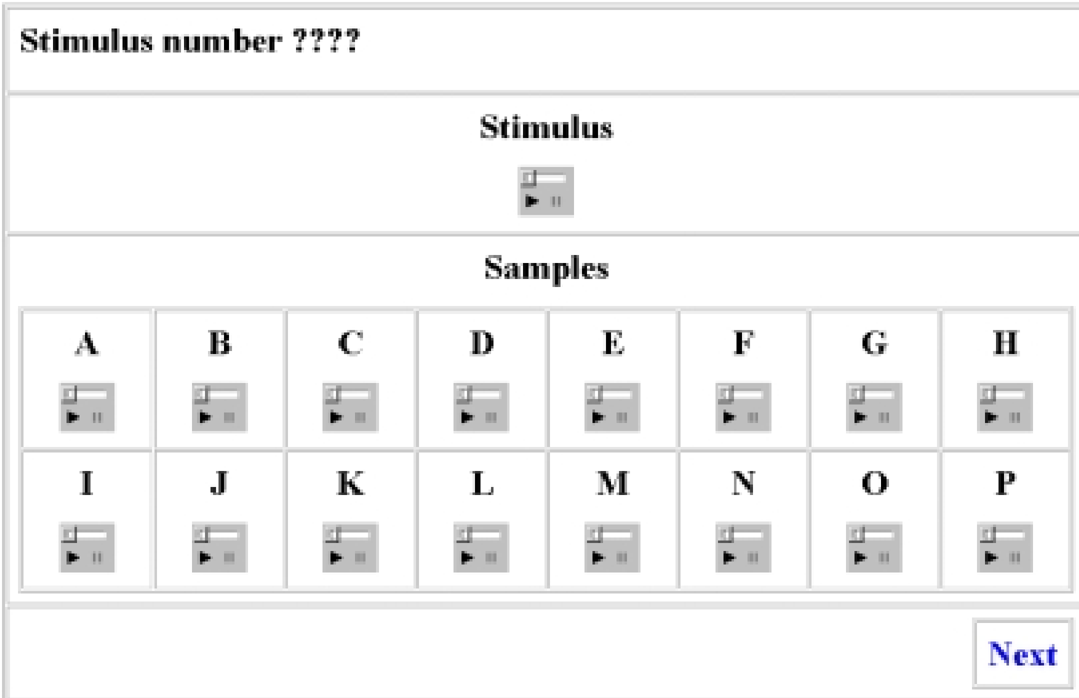
However you perceive the position and dimensions of the sound – either similar to one of the given diagrams or not, you are asked to judge the maximum dimensions in the horizontal plane (width and depth), as shown below.



You should then select one out of the sixteen samples that you think most closely matches the stimulus based on these spatial attributes.

If you find that you perceive the dimensions of the stimulus to be beyond the range of the samples given (either smaller or larger), please select the closest sample (A if it is smaller and P if it is larger). If you think that the dimensions of the stimulus are very similar to more than one sample, or are between two of the samples, then select the one that you think is closest. Please only select one sample for each stimulus.

The audio will be reproduced over headphones and is controlled using the computer. An example of the user interface is shown below.



Each sound is 5 seconds long. Clicking on any of the play buttons (▶) will play the whole 5 seconds. Clicking on the play button of another sound will start the new sound. If a new play button is clicked while there is a sound playing, the new one will play and the old one will be silenced. You can listen to as many of the sounds and as often as you feel you need.

When you have made a decision, mark your choice in the relevant box on the answer sheet with a cross (as shown below). If you make a mistake, scribble out your cross and mark a new cross in the correct box.

Stimulus No. ????

Please select one of the 16 samples that you think has the most similar maximum size as the stimulus and mark it with a cross

A	B	C	D	E ✕	F	G	H
I	J	K	L	M	N	O	P

Once you have made your choice, click ‘Next’ in the bottom right corner of the screen to move on to the next stimulus. There are 18 to judge in all.

It is important to remember that there is no right or wrong answer. Your judgement is the correct answer and you will not be marked on your choices. You will not be told anything about the sounds involved or the processing carried out, if any. If you are interested, you can find out after all the listening tests have been carried out.

To repeat: I am looking for you to match each stimulus to one of the 16 samples that you think has the most similar maximum size. To do this, match the narrow-band sound to the whole frequency range of the wide-band sound. You should mark the appropriate box on the response sheet with a cross and move on to the next stimulus. There are 18 stimuli in all. You have as much time as you need and can listen to each sound as often as you need. Please familiarise yourself with the samples before the test begins and try to determine any spatial differences you can perceive between them.

Please feel free to ask any questions, preferably before the test begins.

Enjoy the listening, and thank you for taking part.

REFERENCES

[1] Blauert, J. 1972: 'On the Lag of Lateralization Caused by Interaural Time and Intensity Differences', **Audiology**, vol. 11, pp. 265-270.

[2] Griesinger, D. 1997: 'The Psychoacoustics of Apparent Source Width, Spaciousness and Envelopment in Performance Spaces', **Acta Acustica**, vol. 83, no. 4 (July/August), pp. 721-731.

[3] Grantham, W. D. and Wightman, F. L. 1978: 'Detectability of varying interaural temporal differences', **Journal of the Acoustical Society of America**, vol. 63, no. 2 (February), pp. 511-523.

[4] Griesinger, D. 1992: 'Measures of Spatial Impression and Reverberance Based on the Physiology of Human Hearing', **Proceedings of the 11th International Audio Engineering Society Conference**, Portland, Oregon, USA, pp. 114-145.

[5] Griesinger, D. 1998: 'General Overview of Spatial Impression, Envelopment, Localization, and Externalization', **Proceedings of the 15th International Audio Engineering Society Conference**, Copenhagen, Denmark, pp. 136-149.

[6] Mason, R. and Rumsey, F. 2000: 'An Assessment of the Spatial Performance of Virtual Home Theatre Algorithms by Subjective and Objective Methods', **Audio Engineering Society Preprint**, 108th Convention, preprint no. 5137.

[7] Mason, R., Rumsey, F. and de Bruyn, B. 2000: 'An investigation of interaural time difference fluctuations, part 1: the subjective spatial effect of fluctuations delivered over headphones', **Audio Engineering Society Preprint**, 110th Convention.

[8] Mason R., Ford N., Rumsey F. and de Bruyn B. 2000. 'Verbal and Non-Verbal Elicitation Techniques in the Subjective Assessment of Spatial Sound Reproduction', **Audio Engineering Society Preprint**, 109th Convention, preprint no. 5225.

[9] Bech, S. 1990: 'Listening Tests on Loudspeakers: a discussion of experimental procedures and evaluation of the response data', **Proceedings of the 8th International Audio Engineering Society Conference**, Washington D.C., USA, pp.101-115.

[10] Barron, M. and Marshall, A. H. 1981: 'Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure', **Journal of Sound and Vibration**, vol. 77, no. 2, pp.211-232.

[11] Gabrielsson, A., Hagreman, B., Bech-Kristensen, T. and Lundberg, G. 1990: 'Perceived sound quality of reproductions with different frequency responses and sound levels', **Journal of the Acoustical Society of America**, vol. 88, pp. 1359-1366.

[12] Moore, Brian C. J., Glasberg, Brian R. and Baer, Thomas. 1997: 'A Model for the Prediction of Thresholds, Loudness, and Partial Loudness', **Journal of the Audio Engineering Society**, vol. 45, no. 4 (April), pp. 224-240.

[13] ITU-R BS 1116. 1994: 'Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems', **International Telecommunications Union**, Recommendations ITU-R BS 1116, pp. 276-297.

[14] Howell, D. C. 1997: **Statistical Methods for Psychology**, 4th edition, (Belmont, CA: Duxbury Press).