DESIGN AND TESTING OF A GRAPHICAL MAPPING TOOL FOR ANALYZING SPATIAL AUDIO SCENES

JOHN USHER[1], WIESLAW WOSZCZYK[2]

[1] *Multichannel Audio Research Laboratory*
jusher@po-box.mcgill.ca
[2] *Centre for Interdisciplinary Research in Music Media and Technology*
wieslaw@music.mcgill.ca
Graduate program in Sound Recording, McGill University, Montréal, Canada

A computer-driven graphical user interface (GUI) was developed for the evaluation of spatial attributes of simple auditory scenes created using a loudspeaker pair. The GUI allows subjects to draw ellipses to represent locations from which they hear the sound. Anechoic, monophonic recordings of music and speech were used to test the GUI. Overlaid response plots (so called density plots) were used to visualize the spatial hearing of subjects. The sensitivity and reliability of this tool used as the measure of listeners' awareness of the spatial distribution of sound has been investigated and quantified through a series of tests. Results indicate that listeners can reliably and accurately map the audio scene using the system. Different loudspeakers were used to project the sound and the graphical elicitation tool demonstrated differences in spatial distribution of sound between them. A more detailed view of precedence effect is possible with the new tool, and in general, an improved visualization of spatial perception of sound.

INTRODUCTION

The recent increase of interests in spatial audio by researchers and the commercial market has created a demand for new and effective ways to evaluate a listening experience from a spatial perspective. We have developed and evaluated a computer-based graphical-mapping tool, a Graphical User Interface (GUI), to evaluate loudspeaker audio systems. We describe how the tool can be used to represent certain perceptual spatial features of an audio scene. The GUI will be used to design, evaluate, and develop a multi-channel loudspeaker system. In order to do this comprehensively, the software has been constructed to allow the two acoustic components of any acoustic scene to be represented: the direct and the indirect sound (synthetic or natural). The features we investigate are related to the sound image which a listener perceives to be part of the recorded source (e.g. a musical instrument). In this paper, we investigate the application of the GUI for representing the phantom images a listener would hear as direct sound from the recorded source, which we will call "source images" to distinguish them from those images created by the indirect sound. The purpose of this study is to evaluate the GUI within a highly controlled experimental setup with the sound reproduction limited to a loudspeaker pair in front of the listener. In a further paper we will discuss how the GUI can be used to evaluate the perceptual sound images associated with both of these acoustic components for multi-channel loudspeaker configurations.

The task of mapping "where we hear the sound" is not a new one. The resolution and accuracy of an audio spatial map is critical when we are comparing different audio experiences. Other considerations for designing spatial audio mapping tools have been investigated by Mason et al in [1]. A highlighted concern is the tools ease of use, or intuitiveness. This factor encompasses such concepts as ambiguity of language and mental imagery, the latter of which is affected by the abstraction of the map to the listeners "real world"; that is, the tools ego-centricity. A further consideration is the drawing method employed for spatially representing where the sound is heard; different approaches to this task have yielded particular problems. Free-drawing mapping systems (e.g. [2, 3, 4, 5]) offer a simple solution for representing the sound scene, but suffer from the problems of emotive bias, e.g. the listener may draw a jagged image to represent a "jagged" sound, such as a trumpet (as was found in [5]). Furthermore, it can be difficult to statistically interpret freely-drawn shapes, e.g. to find the geometric centre of the object. We have chosen to use ellipses for representing regions of space the listener hears the source image(s), and we believe that using this constrained mapping system will help to eliminate some of the problems involved with using free-drawing systems.

To describe the homogeneity of the sound scene, we introduce the idea of "image hot-spots". These can be thought of as local regions of certainty about the listeners perception of a sound image at a particular location. Some listeners use the term "image density" to describe the sound distribution within a phantom-image, where a dense image region would correspond to a local hot-spot. This idea of "spatial certainty" has been used before in audio sound imaging work [6, 7], and we use the analogy of image "hotness" to describe the concept to users of the GUI. The term "hot-spot" was chosen as it is familiar to participants in the experiments, who undergo a critical spatial listening program at McGill University and often observe that individual sound images are undesirably spatially distributed within an audio scene.

1. DESIGN OF THE GUI

The tool we describe is designed to provide a simple and intuitive interface to allow the user to describe a sound scene using two variables: space and image density (i.e. "hotness" or certainty). Therefore, the GUI is not so much a sound-scene analyser as a perceptual data collector and is the "front end" of the spatial analysis system.
Implicit in this kind of tool is that the evaluation is of the sound *character* rather than *quality*, a distinction noted and described by Rumsey [8]. That is to say, this elicitation tool records a descriptive spatial judgement rather than an emotive-driven preference rating.
The level of abstraction which a person uses to describe a percept is generally a trade-off between how meaningful and relevant the personal description is to the individual concerned, and how well this description can be understood and interpreted by others. Various techniques deal with this problem. Two general categories of language have been used for spatially describing auditory scenes: a verbal (i.e. semantic) language (e.g. [9]), and a non-verbal descriptor (e.g. [2, 3, 4, 10, 11]). A review of verbal and non-verbal elicitation considerations can be found in [1]. We have chosen a mapping system whereby both the position and extent of the auditory event are described directly, a technique which has been employed with sound localization tasks for at least 40 years [12].

The use of visual cues in auditory spatial perception investigations is a rather contentious issue. The bias of our auditory localization system towards conclusions of our visual localization system is a well documented phenomenon (e.g. [13, 14]). This "intersensory bias" can be so prevalent that our auditory system may indeed re-map its internalized system which translates acoustic localization cues to perceived source location [15]. In the experiments to test our GUI, the loudspeakers are therefore concealed from the listeners behind an acoustically transparent curtain, but with visual references at 10 degree intervals to help the listeners' spatial correspondence between the real world and the GUI map. We believe that this 10° interval is small enough so that any auditory localization bias introduced by the visual interaction will be comparable with the inherent localization errors associated with phantom-image localization, which varies from about 3.5° to 10° [6, 16]. The markers on both the curtain and GUI are coloured to ease correspondence.

Although ego-centric views are very useful for sound localization tasks, in that they provide a very intuitive mapping perspective, it is difficult to describe the distance dimension with them. In order to allow a full lateral spatial evaluation of the audio scene to be conducted by the subjects, we have opted for an exocentric view of the listening environment. This "top-down" plan view has been used in similar experiments (e.g. [4, 6]), and along with the visual reference points we have found that listeners have no problem mapping where the sound-images are heard to the GUI.

1.1. Pilot experiment

1.1.1. Method

We conducted a test experiment to investigate if just a single ellipse could be used to describe where subjects hear the sound with a simple audio set-up. Three subjects took part, all of whom had experience in the task. The sound was presented under blind-listening conditions by a loudspeaker pair at ±30°, and the listeners had to map the sound to a plan-view diagram of their environment with a GUI, similar to that shown in figure 10. The GUI was created using *MATLAB*. Three anechoicly recorded monophonic stimuli were used: speech, bongos, and a double bass. The recordings were presented at 3 sound levels: 57, 63, and 69 dB_A and each of the 9 unique permutations was repeated 4 or 5 times. The 9 different stimuli were presented in random order. Listeners could freely rotate their heads, although they were asked to visually line up two suspended markers to ensure their heads were at equal distance to both loudspeakers.

1.1.2. Results

For each of the permutations, the elicited responses were overlayed so we can see where responses spatially overlap. These overlayed plots ("density plots") can be used to describe how consistently the same person represents the same audio scene using the GUI (a within-subject comparison), or how consistently different people represent the same audio scene (a between-subjects comparison). This technique has been used for at least 30 years to investigate spatial perception of audio, whereby the super-positioning of the responses may be accomplished using photography [2] or computers [4]. The raw density plots from this pilot experiment can be seen in figure 9. We will see later how further processing of these plots renders a more intuitive and readily interpreted presentation of where and how the sound is heard.
We calculated the degree to which responses overlapped with a measure devised by Mason in[1] (used in experiments in [4]). This "Similarity statistic" (S) is a value between 1 and 0 which is proportional to the percentage of over-lapped response area (1 is 100% overlapped). We assumed that each ellipse has a weight of 1 (or "density" of 1), and will linearly sum with any other ellipse which occupies the same area when overlayed.

$$S = \frac{\sum_{n=1}^{N}(n-1)A_n}{(N-1)\sum_{n=1}^{N}A_n} \qquad (1)$$

where:
A_n = area of response with a density level of n
N = number of summed response sheets

We found the within-subject consistency to be remarkably high, as can be shown in figure 1, which gave a value for S of 0.834. In this pilot study, the mean within-subject similarity statistic was 0.276, and the between-subject value 0.102 The between-subject similarity was typically much higher than that found in a similar spatial audio scene mapping experiment using a loudspeaker pair and a phase-modulated noise source ([4], chapter 5), for which S ranged from 0.018 to 0.060 (mean value 0.041).

1.1.3. Discussion

We found that using even a single ellipse provided us with data wherein clear trends could be identified by visual inspection of the density plots. Such a trend was the proportional relationship between image (i.e. ellipse) width and loudness. We might intuitively hypothesise that louder sounds are perceived as spatially larger, a form of perceptual auditory constancy, and we can confirm this

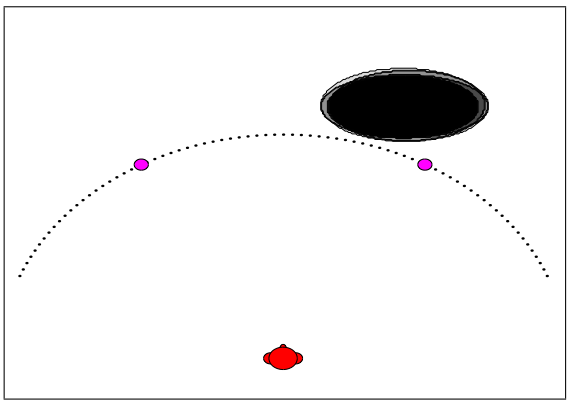

Figure 1: Density plot from subject C for bass stimulus presented at 63 dB_A, summed from 4 repeated random presentations. S=0.83

from both visual and statistical inspection of the density plots in figure 9. We investigated this in terms of the relative and absolute image widths, i.e. in degrees and metres (respectively).
The results are summarised in figure 2.
The trend in increasing width with loudness is statististically significant for both the speech and bongo stimuli *[F(2,39) = 4.235, p=0.022]* and *[F(2,39) = 8.444, p=0.001]* respectively, and for the bass stimuli the trend might be said to be statistically "interesting", *[F(2,39) = 2.674, p=0.082]*. We can also clearly see that the speech sound images are perceived to be generally wider than the others, which we will discuss later.

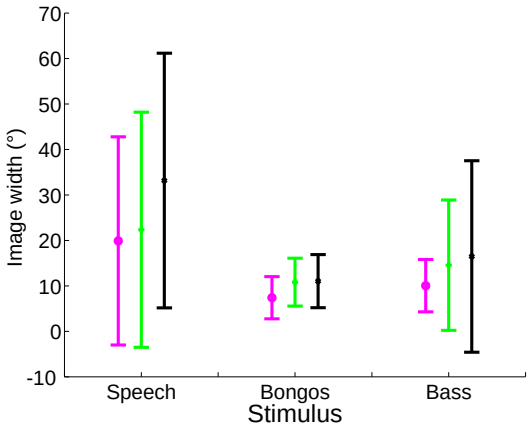

Figure 2: Perceived relative image widths in pilot experiment from 3 subjects. Mean and 95% CL's shown. Data grouped by stimulus, and ordered by loudness: 57, 63, and 69 dB_A.

1.1.4. Conclusion

The results from this pilot experiment demonstrate that the GUI presents a means to reliably elicit a basic spatial envelope for single sound images. We find the high response similarity for representing a repeated sound presentation, as shown in figure 1, shows that trained users of the GUI can be expected to consistently elicit where they hear the sound images in a simple audio scene. This applied to both the within-subject and between-subject density plots, wherein the similarity statistic far-exceeded that of similar experiments using a free-hand drawing system. It might be argued that the subjects simply remembered where they heard that particular instrument sound image from a previous presentation. However, if this was the case then we would not see the statistical trends for the independent variable loudness as we do. Furthermore, if we compare the GUI plots of presentations for different loudness levels, we do not see such a high response consistency.

A common comment from users of the GUI was of frustration at not always representing the full spatial scene. This was on account of the inherent limitations of using a single ellipse to temporally average where the sound was heard. We therefore extend the mapping options available to the GUI user to include multiple ellipses, and ellipse weighting to describe the sound distribution within these images.

To describe a non-uniform distribution of sound within a single image, local regions of differentiation can be described using so called "hot-spots". These may be used in the case of a temporally-shifting image, whereby images may appear for brief periods of time at roughly defined locations. Although the GUI will not be able to represent the temporal detail of the images, the listener can grade their "spatial certainty" for these local regions using a weighting system analogous to a hot-cold temperature scale, with "hot" corresponding to certain. The subjects who will use the GUI in the next experiment are familiar with this concept from academic experience in critical spatial listening classes. We believe that with the addition of this functionality and the ability to draw multiple image regions, the GUI will provide a means to accurately describe where and how the sound is heard, and will provide us with results which can be explained with existing data and theory concerning how we spatially hear phantom-images in loudspeaker audio scenes.

2. TESTING THE GUI

We will now describe an experiment which uses the GUI to measure how we hear time-panned sound images in a 2-speaker blind listening test. The GUI is designed to evaluate multiple-speaker audio systems (typically 5-channel, arranged according to the conventional 5-speaker configuration as in ITU-R BS 775 [17]). However, in this initial study we are interested in evaluating the performance of the GUI, not the audio system, so we limit the set-up to a two speaker arrangement with the speakers at $\pm 30°$. Furthermore, here we are concerned with the elicitation of the source images only, and much research has been undertaken on this using a two speaker paradigm (see [16] chapter 3 for a review). Thus we can evaluate the accuracy of data from the GUI using existing data from these experiments.

We have designed a simple experiment to evaluate how sensitive the GUI is for representing where listeners hear sound in a two speaker sound scene; a sound is played from one speaker and a delayed copy at the other. We will call the two loudspeakers the leading and lagging speaker (i.e. the delayed sound radiates from the lagging speaker), and we will use a range of inter-speaker delays in the experiment. Small time delays should cause a shift in the image location from the centre of the audio scene to the leading speaker, which is often called "time (delay) panning". Existing research has shown that larger time delays cause the fused image to separate into images centred over each speaker. The delays associated with time-panning are under 1 ms, when our auditory system perceives a shift in image location towards the leading speaker if we increase the inter-speaker time delay. This process has been called the "law of the first wavefront" [16] and the precedence effect. For speaker delays greater than this time boundary where the law of the first wavefront applies, the presence of the lagging speaker might be obvious to the listener. We will discuss the cognitive implications for understanding this audio phenomenon later.

2.1. Method

2.1.1. Stimuli

Three different recorded sounds were used in this experiment: Speech (Danish), bongos (from [18]), and double-bass. All were anechoicly recorded with a single microphone. Each motif lasted about 3 minutes, and would then repeat until the user was ready for the next presentation. The presentation of the stimuli was controlled by a program running on the same computer as the GUI, and created a random-ordered playlist for each subject. The 30 unique audio stimuli were presented once to each subject, though a break of up to 5 minutes was advised after 15 stimuli. The tests typically took 30 minutes to complete.
The sound pressure level at the listening position was measured to be approximately 57 dB_A.

2.1.2. Set-up

As with the pilot experiment, this experiment was conducted in an acoustically damped room (reverberation time at 125 Hz = 0.27 seconds, 250 Hz and 500 Hz RT = 0.17 s). The loudspeakers were hidden behind the curtain, which has 11 markings as described and shown in figure 10. An electro-acoustically matched pair of Beolab 4000 speakers (manufactured by Bang and Olufsen) positioned 30° to the left and right of the listener (i.e. a base angle of 60°) at a distance of 2.3 m (just behind the curtain). 8 subjects took part in the experiment, all members of the sound-recording department at McGill University. These participants were informed that there could be any number of loudspeaker being used in the experiment, and that they could be located anywhere behind the curtain. The given instructions were simple; to draw ellipses with a mouse using the GUI to represent where they heard the sound images, and to weight these ellipses according to how sure they were about hearing sound at that location. 10 training presentations were given to help familiarise the subjects with the GUI. Ten inter-speaker time delays were used: 0, 0.2, 0.4, 0.6, 1, 2, 10, 20, 40, and 60 milliseconds. These delays were created using a Yamaha 03D mixer with an audio sample-rate of 44.1 kHz, therefore we could only have delay times at integer factors of the sampling interval. We measured the electronic latency of the left/ right speaker outputs using a B&K type 2035 signal analyser with a time resolution of 7.81 μs, and found the 10 delays to be: 0, 0.2286, 0.4305, 0.6134, 1.0211, 1.2268, 10.0335, 20.0889, 40.0405, and 60.0240 ms. We also measured these acoustic latencies at the listening position using a single microphone and a dummy-head, and found the speaker acoustic delays to be within 5% of the electronic delays.
As with the pilot, in this experiment the listeners were free to rotate their heads. However, a head-movement of only 3 cm left or right could cause an inter-aural time difference of 0.2 ms, so the listeners were told to keep their heads aligned with two visual markers; the central marker on the curtain and another thread directly in front.

2.2. Results

Figures 11 to 13 show density the plots from this experiment, grouped by stimuli and ordered by inter-speaker delay. A great deal of information is held within these density plots, so we have processed this raw data to display the information in a more readily accessible style from which the sensitivity of the GUI can be evaluated as much by eye as with statistics. This was done by summing the density plots as a function of azimuth, so we can see the weighted extent of the perceived sound for different inter-speaker delays. We present these polar "image directivity plots" overlayed on the density plots. This is similar to loudspeaker output or microphone sensitivity polar plots, where we can see at a glance a vector representation of where sound is heard, showing the lateral angle and the magnitude of image density at this angle. We further develop the presentation of these directivity polar plots to display the image directivity trends as a function of inter-speaker delay. These image directivity plots are shown in figures 6 to 8.
In figure 3 we show how the total summed ellipse areas change as a function of time delay. As we are summing the GUI responses from all 8 subjects, we are really looking at the density-plot volume. The elicited response area is relatively constant for delays between 0.2 and 2 ms, but somewhere between 2 and 10 ms, the area increase more dramatically. Figure 4 shows us where this additional area "comes from"; it comes from increasing image size (and indeed weight, as figures 11 to 13 show) over the right-hand speaker, i.e. the delayed sound becomes more audible at the lagging speaker.
We also see from figure 5 that the listeners are using more regions to describe their audio scene at higher inter-speaker delays; approximately 1 for low delays (i.e. under 2 ms), two regions for delays between 2 and 20 ms, and 3 for delays greater than 20 ms. The number of "regions" in the audio spatial map, is not the same as the number of ellipses. This is simply because if we wish to describe a non-uniform distribution of sound within an image, we must use multiple ellipses to repre-

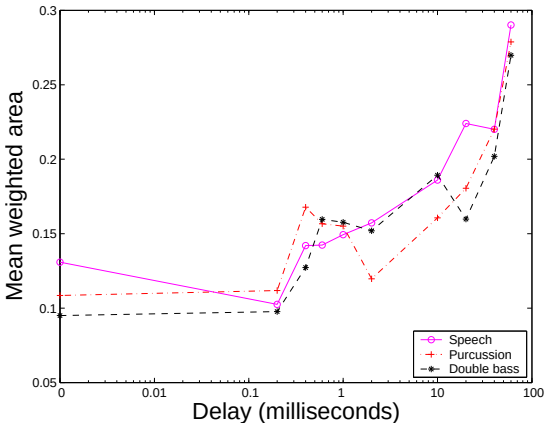

Figure 3: Total weighted area (i.e. "volume") of density plots.

sent the "hot-spots"; there will therefore be more ellipses than regions.

There seems to be a certain inter-speaker time delay at which our auditory system decides that the two sounds it hears from different spatial locations have originated from different sources. Users of the GUI demonstrate this by drawing a sound image over the lagging speaker, and assigning it a "hot" weighting. This can be seen by the lobes on the image-directivity polar plots and the appearance of an image at +30° in figures 6 to 8. In our experiment, the GUI shows this time is approximately 10 ms. We will call this time the fusion boundary, as it corresponds to a delay up to which the auditory system fuses both the leading and lagging sound into a single image. We find it very interesting that before this fusion boundary, between 2 and 10 ms, the area for all three stimuli is at a local minima. Furthermore, for this speaker-delay interval we can see from the image polar plots that the single image perceived over the leading speaker is very well defined, with a small width and high density.

It appears that the audio presented with an interspeaker delay of 0.2 ms did not greatly affect the overall perceived image size, as figure 3 shows, but the images were pulled slightly to the leading (i.e. left) speaker, as shown by the lobes in the polar image directivity plots in figures 11 to 13. That the GUI can show these trends is a positive indicator of its sensitivity for use as a tool to accurately map spatial sound-images.

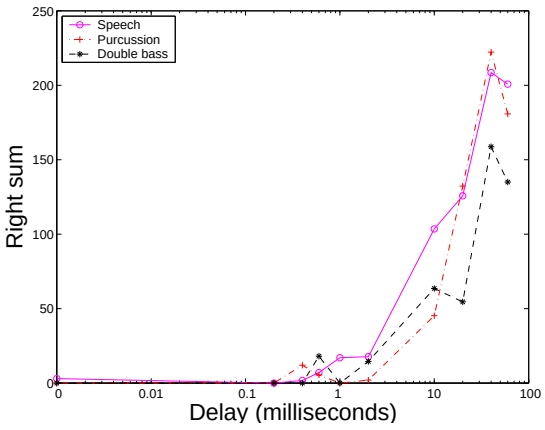

Figure 4: Integrand from +25° to +35° of the density plots, i.e. in the direction of the right speaker.

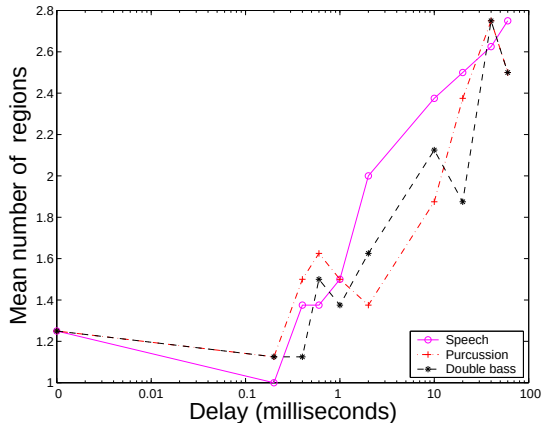

Figure 5: Number of unique regions used to describe the sound images.

Figures 6 to 8 show linear weighted "image directivity" plots as a function of inter-speaker delay.

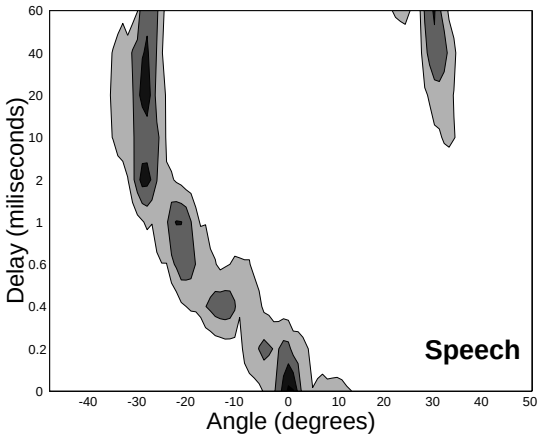

Figure 6:

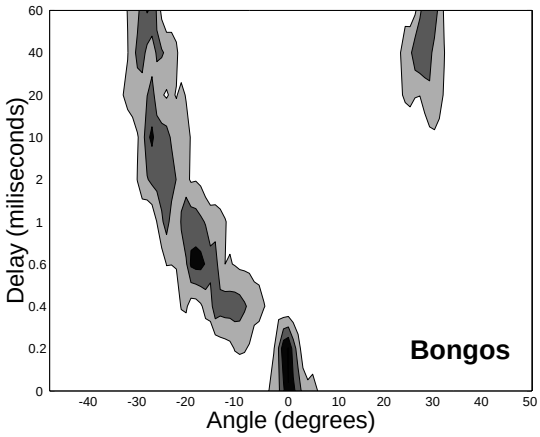

Figure 7:

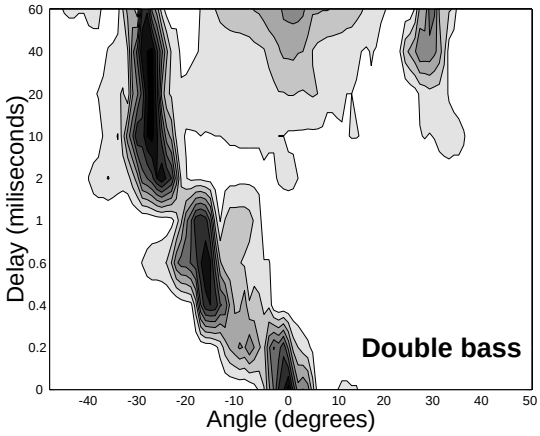

Figure 8:

2.3. Discussion

We will now discuss why we believe the data collected by the GUI can be trusted as an accurate confirmation of a spatial audio experience. If we can provide a cognitive hypothesis for how the sound images should be perceived in the experiment we have described, then the accuracy of the GUI can be measured by how well it provides results which conform to this hypothesis.

Let us consider a two speaker scenario arranged as in our experiment, whereby a sound is reproduced at the left speaker, and a delayed copy at the right speaker.
We posit that the brain interprets the sound from the lagging loudspeaker as a reflection of a sound source located at the leading loudspeaker. Now, if this was the case then it would take a time T_{S-S} seconds for sound to travel from the source (i.e. the leading speaker) to the reflecting object (lagging speaker):

$$T_{s-s} = \frac{2(ds \cdot \sin \frac{\alpha}{2})}{c} \qquad (2)$$

where:
c = speed of sound
ds = listener → speaker distance
α = Loudspeaker "Base angle"
In the loudspeaker environment, we presume the brain can localize each speaker accurately using "hard-wired" sound localizing mechanisms in the peripheral auditory system (i.e. the process which the Jeffress-model describes, see [16] for details). Using the auditory systems learnt (we assume) knowledge of acoustical laws, the brain calculates that if the sound coming from the lagging speaker is indeed just a reflection of the sound coming from the leading speaker, not only must it sound similar (which of course it does), but it must take the correct time to travel to this "reflecting surface"; this time is T_{S-S}. So when the interspeaker time delay *is* equal to T_{S-S}, it confirms the brains hypothesis that the two sounds have a similar origin and will therefore have no reason to perceptually segregate them.

Now, let us look at our experimental paradigm where α is 60° and ds is 2.3 m. Using equation 2 we get a value for T_{S-S} of 6.7 ms. As mentioned, we can see that the total ellipse area is a local minimum in the region of 2-10 ms (see figure 3), and we can see that for speaker delays above about 10 ms we do not fuse both the leading and lagging sounds into a single image. We can see this trend from visual inspection of both the raw density

plots (figures 11 to 13) and the image-directivity plots (figures 6 to 8). In the image-directivity plots, we can see a region of high image density between 2 and 10 ms (i.e. the darker regions). Furthermore, we also find that in other experiments that for a certain inter-speaker delay, the listener is less sensitive to single echoes than for delay times locally before or after. A good example can be found in a study by Olive and Toole on perceptual detection of single sound reflections [19]. They conducted a very similar experiment to ours with a sound coming from one speaker, and a delayed copy from another; a speaker pair with a base-angle of 65°, at a distance of 2m. For their experimental set-up, using equation 2 T_{S-S} is 6.3 ms. They found that the absolute threshold for listeners to detect the delayed sound was highest for sound delays at approximately 7 and 10 ms for recorded speech and castanet stimuli. In other words, it seems from both their and our experiment that the auditory system best fuses the leading and lagging sound when it comes from the direction of the lagging speaker at the time it *should* if there was a single real source at the leading speaker, and a reflecting object at the lagging speaker.
As we did not include a measure for an inter-speaker delay between 2 and 10 ms, we can not precisely see the GUI response at the predicted T_{S-S} value of 6.7 ms. More research needs to be conducted to evaluate the robustness of our cognitive model for spatial imaging using time-panning, such as different speaker base-angles and using time-delay values centred about the T_{s-s} value.

We are aware of the caution which must be applied when analysing the density plots from the processed GUI data. An example of an as-yet unresolved problem is the issue of scaling or weighting the image "hot-spots". We have assigned an arbitrary linear weighting to the 5 hot-spot "temperatures" (i.e. 1 to 5). Therefore, when we layer the individual responses to obtain the density plots, this scaling would imply we hear a hot-spot with a temperature of 2 (i.e. "Cooler than average") with twice as much spatial certainty as a region of the audio scene which we assign a temperature of 1 (i.e. "Very uncertain"). We can not assume the scale for spatial certainty is interval; it is most likely ordinal. Using a non-linear factor such as a logarithmic scaler could enable us to compress the image "temperature" range, and to see when an experimental variable (such as the inter-speaker delay) starts to affect our spatial image perception.

We found in both experiments that the sound images elicited from the speech stimuli were larger than images for the bongos or bass. We believe this is a consequence of the complex spectral and temporal structure of the speech stimuli; it contains both sustained sounds (vowels) and transient consonants. The sustained tonal sounds radiated from each speaker will cause a change in the sound timbre at the ear due to their acoustic interaction. As we are all familiar with the tonal quality of speech-like sounds, the ear may be more sensitive to a timbral change in speech than for the other sounds, which may cause the leading and lagging sounds to fuse less completely; with the result that we hear a wider sound image. We would expect the presence of the lagging speaker to be revealed easier if it was not perfectly phase-aligned with the leading speaker, e.g. due to impedance or acoustic loading mismatches. This would cause a spectral misalignment of the sounds arriving at each ear from the speakers, as well as a non-constant group delay. Therefore, the brain might conclude the two sounds originate from different sources, and would not perceptually fuse them to form a single percept (i.e. a single, well defined sound image). We might expect the motifs with a more transient nature to reveal these phase-misalignments, such as the percussion and speech. Indeed, there is evidence that the localization blur (or Minimum Audible Angle) for real sources can vary considerably for stimuli of different frequency content, generally increasing blur with frequency as in the case of gaussian tone bursts [20]. This could explain why the speech image in the pilot experiment was generally perceived to be wider (though this could be because the speech was perceptually louder than the other stimuli).

In both of our experiments, subjects unanimously agreed that image widths were larger than image depths. We believe this is because of the lack of source-distance cues in the anechoic recordings (i.e. early reflections). However, as this tool is primarily concerned with measuring the spatial distribution of sound in the azimuthal plane, we are not greatly concerned with this issue. Furthermore, the polar sound image technique we have employed for analysing the density plots integrates across the depth dimension of the lateral plane, so by fixing the elicited image depths to a constant value, we can force the polar imaging measure to be insensitive to variation in image depth. Of course, there are many different spatial goals concerning source-image perception in au-

dio, and an understanding of the relevant acoustical saliences which combine to affect our spatial audio perception will not only speed up the process of achieving these goals, but may reveal the processes by which the brain accomplishes this.

3. CONCLUSION

Two experiments were conducted to evaluate the performance of a computer-driven user-mapping tool for measuring where and how a listener hears sound in a two speaker audio scene. We have used various graphical and statistical methods for analysing the data provided by the mapping tool. The data obtained provides us with results which are consistent with existing data on similar spatial sound localization tasks, and which can be explained using a cognitive model of phantom image localization using inter-speaker time differences.

We are satisfied the tool can be used to represent source-related sound images in the lateral plane for audio presented on a loudspeaker pair, and believe the tool can be used to display these sound images for multi-channel audio systems. A further study will investigate how the tool can be adapted to allow listeners to represent where they hear both the source and room related sound components in complex audio scenes.

4. ACKNOWLEDGEMENTS

We are grateful for financial aid for this project provided by Bang and Olufsen and the Natural Science and Engineering Research Council of Canada. We would also like to thank all who participated and helped in the experiments, especially Tallulah for her sound advice.

References

[1] R. Mason, N. Ford, F. Rumsey, and B. de Bruyn. Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction. In *109th Convention of the Audio Engineering Society*, Los Angeles. Preprint 5225, 2000.

[2] B. Wagener. Räumliche Verteilungen der Hörrichtungen in synthetischen Schallfeldern ("Distributions of hearing directions in synthetic sound fields"). *Acustica*, 25:203–219, 1971.

[3] J. Blauert and W. Lindemann. Spatial mapping of intracranial auditory events for various degrees of interaural coherence. *Journal of the Audio Engineering Society*, 79(3):806–813, 1986.

[4] R. Mason. *Elicitation and measurement of auditory spatial attributes in reproduced sound*. PhD thesis, Univerity of Surrey, England. School of Performing Arts, February 2002.

[5] W. R. Woszczyk. Quality assessment of multichannel sound recordings. In *12th International Conference of the Audio Engineering Society*, pages 197–218, 1993.

[6] J. Corey and W. Woszczyk. Localization of lateral phantom images in a 5-channel system with and without simulated early reflections. In *113th Conference of the Audio Engineering Society*, Los Angeles. Preprint 5673, 2002.

[7] T. Lund. Enhanced localization in 5.1 production. In *109th Convention of the Audio Engineering Society*, Los Angeles. Preprint 5243, 2000.

[8] F. Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, 50(9):652–666, 2002.

[9] J. Berg and F. Rumsey. Spatial attribute identification and scaling by repertory grid technique and other methods. *Proceedings of the 16th International Audio Engineering Society Conference, Rovaniemi, Finland*, pages 51–66, 1999.

[10] J. S. Usher. Computational auditory scene analysis of two channel audio material to predict image locations spectrally, as would be perceived on a loudspeaker pair. Undergraduate honours project. School of Acoustics, University of Salford, England, 2001. `http://www.music.mcgill.ca/~usher/papers/salford/casalist.pdf`.

[11] N. Ford, F. Rumsey, and T. Nind. Subjective evaluation of perceived spatial differences in car audio systems using a graphical assessment language. In *112th Convention of the Audio Engineering Society*, Munich. Preprint 5547, 2002.

[12] N. Guttman. A mapping of binaural click lateralization. *Journal of the Acoustical Society of America*, 34:87–92, 1962.

[13] D. R. Begault. Auditory and non-auditory factors that potentially influence virtual acoustic imagery. In *Proceedings of the Audio Engineering Society 16th international conference on spatial sound reproduction*, Rovaniemi, Finland, 1999.

[14] R. L. Storms. *Auditory-visual cross-modal perception phenomena*. PhD thesis, Naval Postgraduate School, Monterey, California, 1998.

[15] B.G. Shinn-Cunningham, N.I. Durlach, and R.M. Held. Adapting to supernormal auditory localization cues I: Bias and resolution. *Journal of the Acoustical Society of America*, 103(6):3656–3666, 1998.

[16] J. Blauert. *Spatial hearing: The psychophysics of human sound localization*. MIT Press, Cambridge, Mass., revised edition, 1997.

[17] ITU-R. Multichannel stereophonic sound system with and without accompanying picture. Recommendation BS.775-1, International Telecommunication Union Radiocommunication Assembly, 1992-1994 1994.

[18] Bang & Olufsen. Music for archimedes. Compact disc, CD BO 101.

[19] S. E. Olive and F. E. Toole. The detection of reflections in typical rooms. *Journal of the Audio Engineering Society*, 37:539–553, 1989.

[20] A.W. Mills. On the minimum audible angle. *Journal of the Acoustical Society of America*, 30:237–246, 1958.

A. DENSITY PLOTS FROM PILOT EXPERIMENT

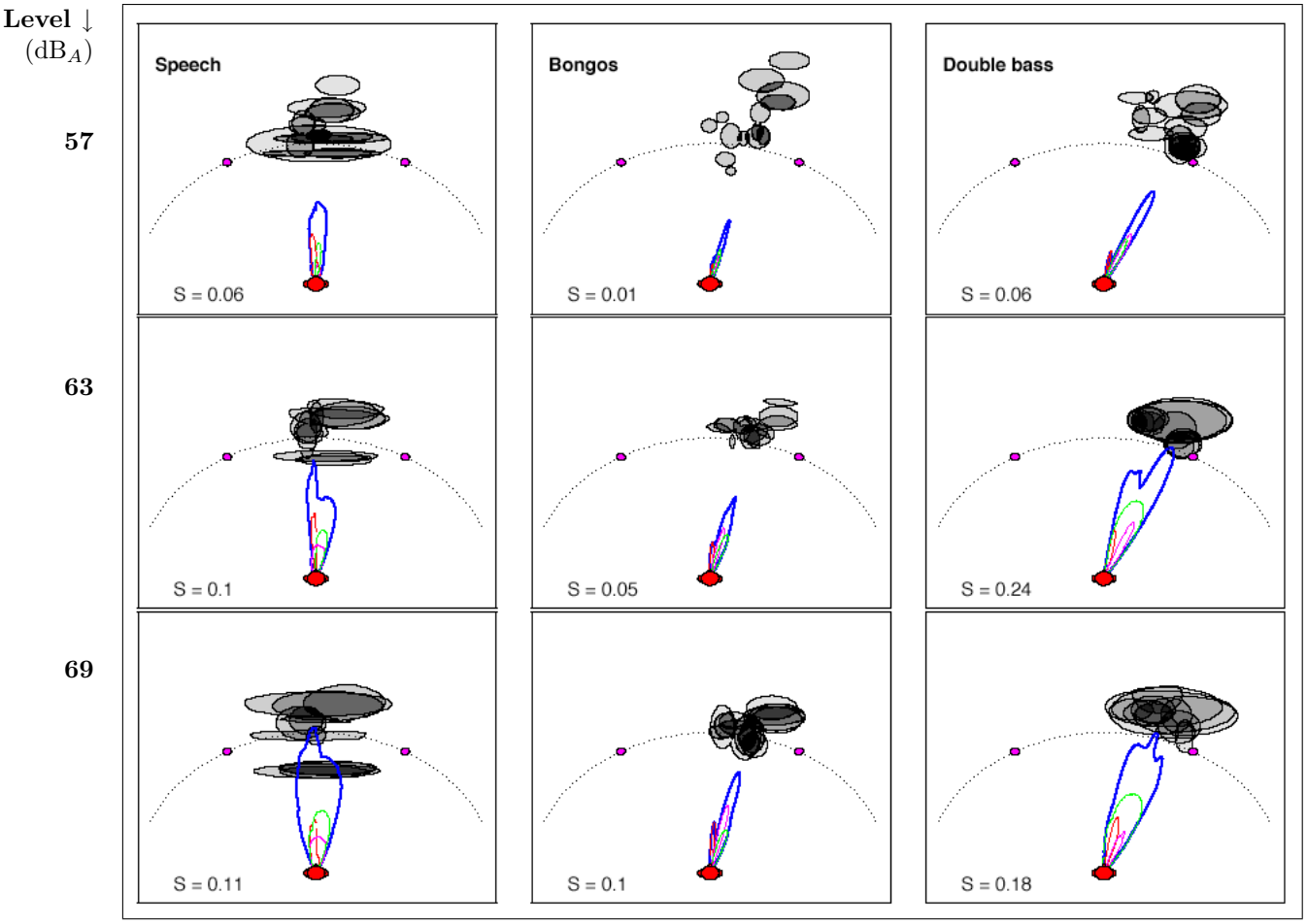

Figure 9: Density plots composed of three summed subject responses for three mono, anechoic stimuli amplitude panned at discrete locations. The subjects could draw only a single ellipse to represent where sound was heard. Contour "image directivity" plot overlayed (thick line), and contribution of each subject shown (thin lines).
Location of the two speakers shown relative to the listener. S = similarity statistic.

B. SCREEN-SHOT OF THE GUI

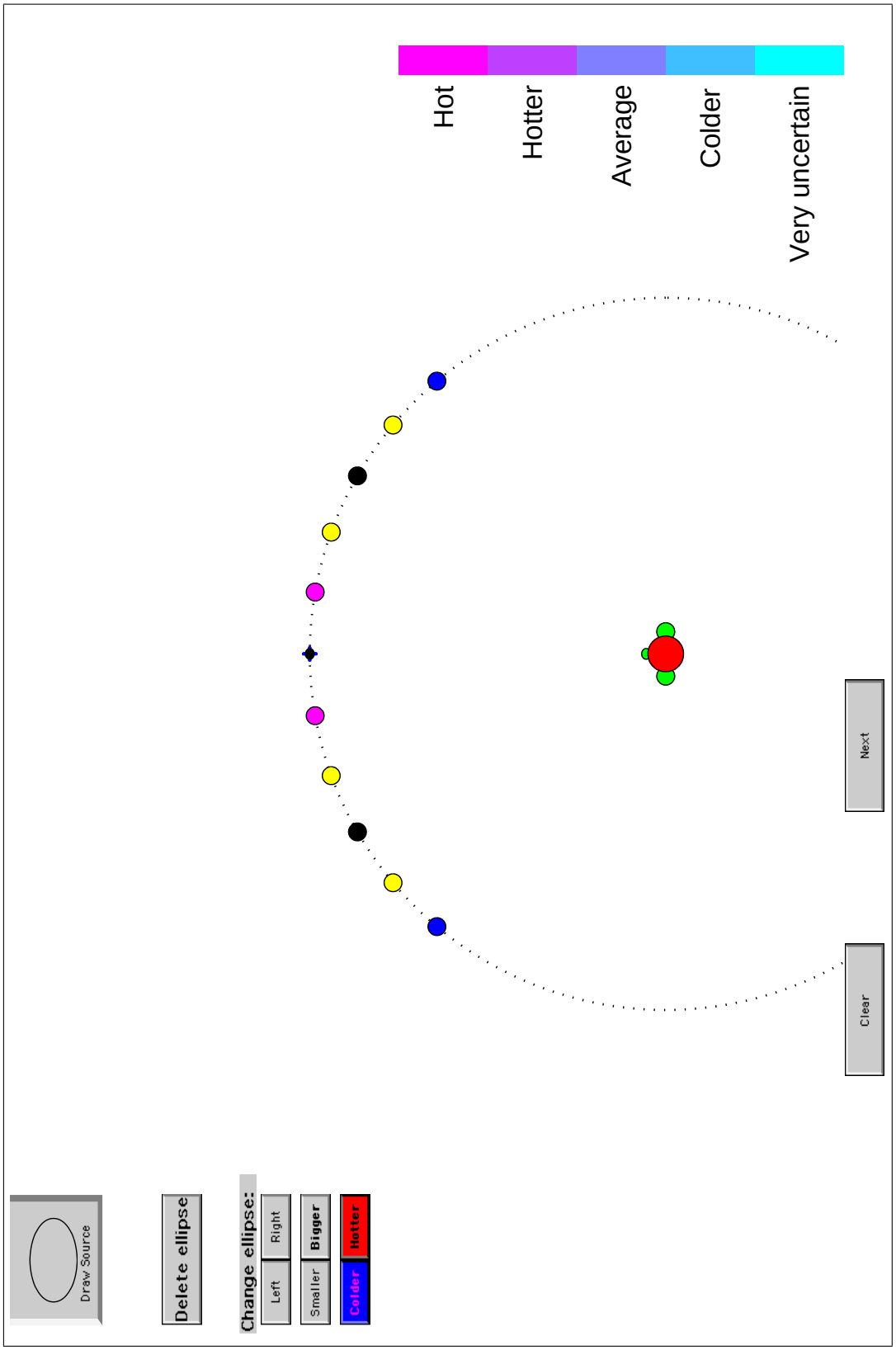

Figure 10: Screen-shot of the GUI as used in delay experiment

C. DENSITY PLOTS FROM DELAY EXPERIMENT

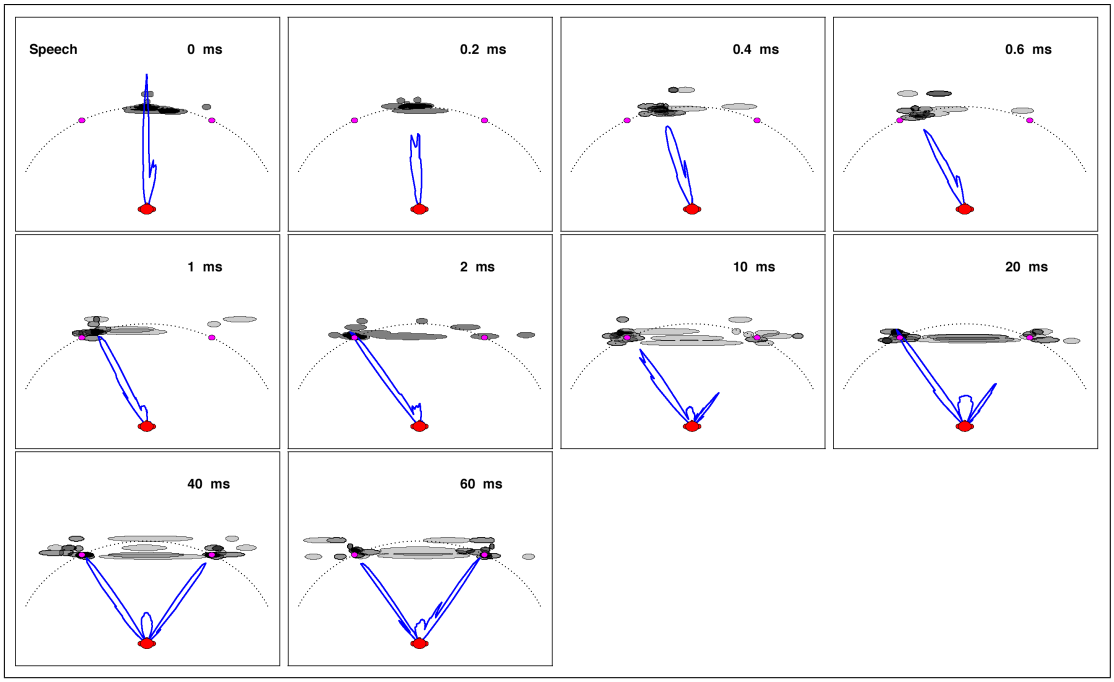

Figure 11: Voice stimuli

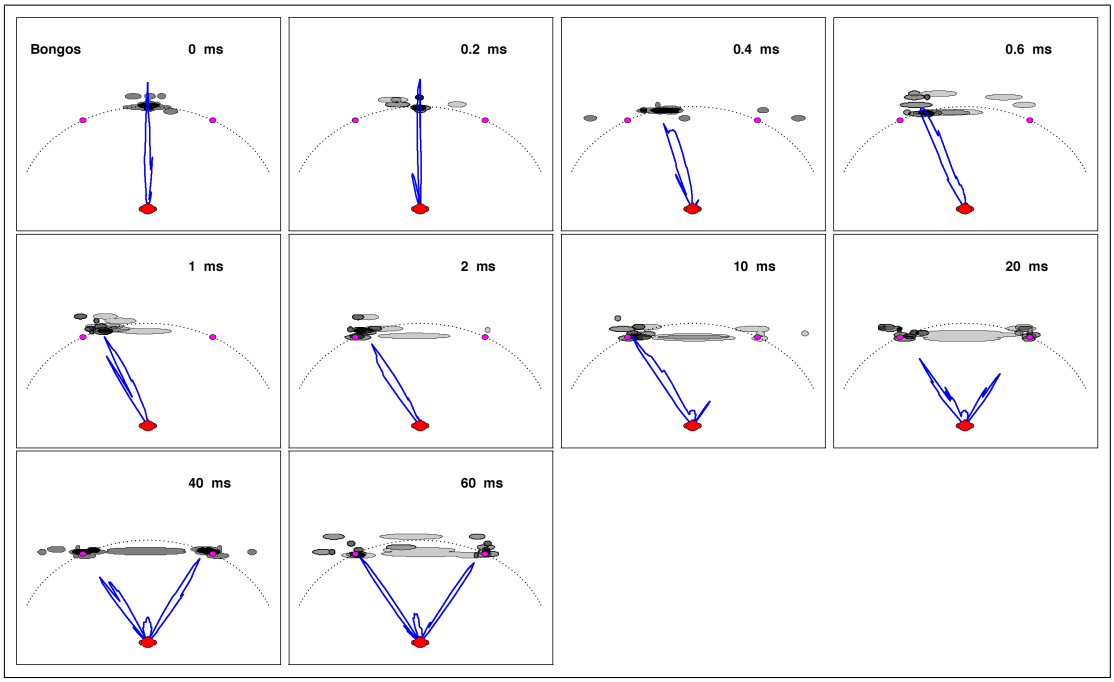

Figure 12: Bongo stimuli

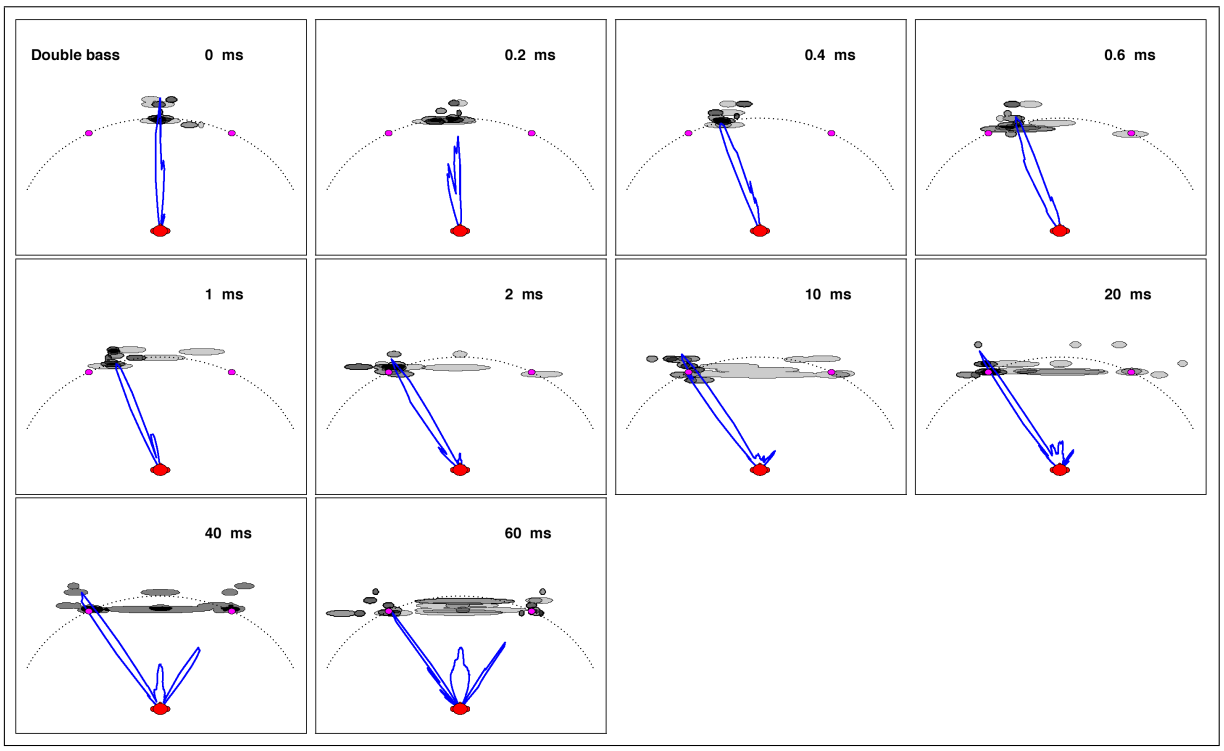

Figure 13: Bass stimuli