TRAJECTORY CAPTURE IN FRONTAL PLANE GEOMETRY FOR VISUALLY IMPAIRED

Martin Talbot and Bill Cowan

David R. Cheriton School of Computer Science
University of Waterloo, Waterloo, Canada
talbotm@acm.org

ABSTRACT

Users who are blind, or whose visual attention is otherwise occupied, can benefit from an auditory representation of their immediate environment. To create it a video camera senses the environment, which is converted into synthetic audio streams that represent objects. What aspects of the audio signal best encode this information? This paper compares four encodings that allow users to perceive the simultaneous motion of several objects.

The comparisons are experimental: subjects hear trajectories of objects moving in a virtual 2D plane, encoded as audio streams with complex frequency spectra, and identify the represented motions. One encoding uses panning for horizontal motion and pitch for vertical motion (the Pratt effect). A second uses best-fit head related transfer functions (HRTFs) to localize stream positions. The third combines the first two, using pitch to redundantly code elevation in a HRTF presentation. Finally, the fourth enhances the third, using best-fit HRTF to 'vertically pan' each audio stream at constant but unique elevations, for superior audio segregation.

The fourth method outperforms the other three according to two measures, the accuracy of subjects' perceptions, and the number of replays needed to achieve those perceptions. With it subjects can perceive up to three different simultaneously-presented motions after minimal practice. The results show that the Pratt effect is a more robust method than HRTF for representing vertical motion, and that, combined with the Pratt effect, vertical panning using a HRTF improves motion perception.

1. INTRODUCTION

A long-time objective of assistive technology is allowing the blind to 'see' with sound. To do so a head-mounted video camera captures images; computer vision software analyses the images, locating salient objects; and sounds representing the objects are created as if they emanate from the position of those objects. Such an interface allows the blind to position objects in the environment by hearing their positions. It also provides eyes-free perception to sighted users whose visual attention is otherwise occupied.

Indeed, such technology is now commercially available. For example, vOICe© vision technology for the totally blind offers live camera views based on sophisticated image-to-sound renderings [1]. This system presents static scenes to its user, and the next step is to tackle the problem of scenes that contain motion. Both ego motion, which occurs when the perceiving subject moves, and object motion, which occurs when an object in the scene moves, provide important perceptual information to the perceiver [2].

Motion defines a trajectory, the path an object follows. To define an object's trajectory from video capture many difficult problems, including object segregation, depth perception and motion capture, must be solved. However, these challenges, which belong to computer vision, lie in the pre-processing stage, which is beyond the present work, which considers a different challenge, how to encode a trajectory for optimal perception.

For locating streams of sound we use the head related transfer function (HRTF) and the Pratt effect (PE). The HRTF encapsulates complex interactions, such as delays, resonances and diffraction, that occur when incoming sound waves interact with the perceiver's head, shoulders, torso and pinnae, before reaching the ear canal [3]. Audio signals presented by headphones require sound pre-filtered by an appropriate HRTF for the perceiver to determine its spatial origin. HRTFs vary from one individual to another, so that we must have the right HRTF for each individual. Measurement of individual HRTFs is prohibitively expensive, but choosing the best fitting HRTF from an existing set, as done here, is an effective low cost practice.

The Pratt effect originates in associations, developed early in life, between elevation and pitch: short wavelength sound being described as high, long wavelength sound as low. Perceivers automatically use these associations when locating the elevation of sounds: high frequency sound is perceived as originating from high elevations, low frequency sound from low elevations [4].

Our experiment uses combinations and modifications of HRTFs and the PE to encode location. Moving sounds, so encoded, are presented to observers, who identify their trajectories. Fast and correct identification is the measure of encoding quality.

The following section describes previous research examining observers' ability to determine location from sound. Then section 3 develops the methodology of our experiment. Section 4 describes the experiment, followed by results and analysis in section 5. The final section discusses the results and future research.

2. PREVIOUS WORK

Sound localization research has a long history. For example, the effects of interaural delays and intensity differences were studied psychophysically in the early 20th century [5]. Batteau [6] recognized that the pinnae assist in spatial localization by filtering audio input, an effect reproduced by the HRTF. Subsequently, HRTF-related localization has been widely studied [3] [7], including its virtual reproduction [8] [9] [10]. Other cues for localization, such as pitch cuing elevation, have also been considered [4]. Sound localization research remains active [3] [11], with ongoing technological innovation based on it. Adaptive computing, which appeared in auditory displays in the 1970s, and electronic travel aid (ETA) systems are good examples.

The first system to help the blind read printed text, the Stereotoner, formed letter shapes as columns of auditory tones. Well-trained in-

dividuals could read sixty words per minute [12]. The earliest auditory display for ETA was the Sonic torch, which used sonar capture and monaural display to help the blind avoid obstacles [13].

These systems were limited by the low dimensionality of their auditory displays. Indeed, encoding many streams of information in sound is hard, stream segregation being difficult in synthetic audio. Consequently, segregation and grouping, which are linked because listeners segregate streams by group stream parts, have been studied for many aspects of sound, including the following.

Timbre: Multidimensional scaling shows correlations in similarities of timbre and audio stream grouping [14] [15] [16].

Pitch: Diana Deutsch showed the segregation of two sets to depend on their frequency separation [17].

Time: Temporal proximity affects segregation: the closer in time sounds are the more they group together [18].

Rhythm: Marie Reiss Jones showed that regular rhythmic patterns segregate better [19].

Space: The American composer Henry Brant described the effect of spatial segregation on orchestration [20].

Here timbre and space are used for segregation. Pitch, which is used to encode motion, is not available; time and rhythm interact with the temporality of encoded motion.

Motion of sound sources has also been studied. The detection and discrimination of simulated horizontal motion was tested by measuring detection and discrimination of 500-Hz tones moved between two fixed loudspeakers [21]. Other researchers investigated the minimum duration time for perceiving auditory apparent motion (MAMA) [21] [22] [23]. Motion induced by auditory cues, such as intensity, binaural delay and the Doppler effect, have been studied experimentally [24]. Motion induced by HRTF, however, seems not to have been examined.

Motion and segregation together have been examined using experiments modelled on ones originally performed on the visual system. Because of the visual heritage in those experiments, stimuli were pure tones, which omit the rich set of features that support auditory segregation. Many commonalities were discovered between vision and audition [25] [26], some of which are exploited in this paper.

None of this research examines how well listeners capture trajectories of simultaneous audio streams. Research of this type is important because the task is important to vision deprived listeners. Also, as described in section 6, it unexpectedly opens up deep questions about the nature of redundant coding in auditory display.

3. METHODOLOGY

This section discusses the general guidelines we adopted, and details of the stimuli, including the model geometry, the HRTF methodology, timbres, and the velocities of the trajectories.

3.1. General Guidelines

The experiment requires observers to perceive one or more trajectories encoded in simultaneous audio streams, which divides their attention. In practice, having simultaneous streams is essential: visual environments usually contain multiple moving objects.

Trajectory recognition is made more difficult by simultaneous streams, owing to interference when trajectories compete for resources in the observer's short-term memory [2]. Thus, the cognitive component of recognition should be as easy as possible: recognition should be categorical, using stimuli that are highly over-learned. Letters and digits fulfil this requirement.

Letters and digits are two-dimensional, naturally represented on a frontal plane perpendicular to the listening direction, which eliminates depth perception. This limitation is accepted because simultaneous trajectories are differentiated by timbre, while auditory distance perception is controlled by loudness and echoic depth. When timbre is complex, loudness interacts with timbre. For example, increasing loudness is not perceived as the inverse of decreasing loudness, even when the changes are symmetric [27]. Echoic distance is similar because timbre interacts with reverberation [28]. Three-dimensional trajectories remain for the future.

Because variations in loudness can be confused with motion in depth loudness is constant. Similarly, the environment is anechoic.

Localization of auditory streams is degraded at low stimulus intensities [29], the effect being greater on localization in elevation than in azimuth. Thus, the intensity throughout the experiment is the intensity of a normal conversation, about 60dB SPL.

Temporal information might help in perceiving trajectories, but this study avoids velocity effects by maintaining velocity along a trajectory within a reasonable range, as discussed in section 3.2.4. Like motion in depth, velocity variation remains for the future.

3.2. Details of the Stimuli

This section discusses characteristics common to our four models. We cover the overall audio resolution of the interfaces (audio pixels), the HRTF method in use, and detail the audio streams' timbral composition.

3.2.1. Frontal plane geometry

Each trajectory lies a plane perpendicular to the listening direction. The origin of the plane is directly in front of the listener. CIPIC©'s sound source locations [30] are specified by azimuth and elevation in interaural-polar coordinates. Simple triangle rules translate these coordinates into planar coordinates. Figure 1 shows the plane with the central point at the origin.

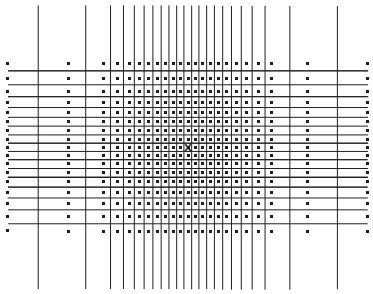

Figure 1: This figure shows the locations of HRTF measurements and the delimiters used for interpolating in the plane. Each grid division is an audio pixel; each dark square is the location of an HRTF measurement. The origin is marked by 'X'.

Resolution degrades in the periphery owing to circular projection. Consequently, CIPIC© data is not uniformly sampled in azimuth, with denser spacing near the midsagittal plane. Thus, the field of view (FoV) is limited to at most [-60°, 60°] in azimuth, even though the database covers a wider range. Furthermore, human motion detection for sound is limited outside a lateral FoV of

[-40°, 40°]: peripheral objects must travel three times as far for motion to be detected [31], a second limitation on the useful FoV.

Because of differential shadowing by the torso, performance differences exist between elevations above and below the horizontal plane, but because the differences are small, the stimuli are centred on the horizontal plane. Transposing the stimuli up, for example, to avoid crossing the horizontal plane, would be unnatural because most audio motion in natural environments lies close to the horizontal plane.

The stimulus plane is 17 vertical audio pixels by 23 horizontal audio pixels. HRTF measurements are not interpolated. Instead delimiters discretize the plane into the audio pixels shown in Figure 1. The Pratt effect uses the HRTF grid delimiter for consistency. Consequently, the four experiments all use the same grid.

3.2.2. HRTF

Scientists began measuring individual HRTFs in the 1970s, placing probe microphones is the ears of human subjects to measure azimuth and elevation dependencies. Unfortunately, perceived elevation has large individual differences so that it is well perceived only using a personally measured HRTF. Three ways of solving this problem exist at present.

1. *Individualized HRTFs*. Each listener's HRTF is measured individually, which gives excellent results, but at very high cost.

2. A *standard or generic HRTF*. Martens *et al.* [11] created a generic HRTF filter, which is convolved with a sound source to simulate spatial images for a general population. However, it gives poor elevation results for many listeners.

3. *The best-fit HRTF*. Each listener is matched to the best HRTF filter in a database of individual HRTF measurements. It requires pre-screening, and often produces satisfactory results. However, the goodness of fit for an individual is hard to assess.

Best-fit HRTFs are used here. They can provide inexact elevation, but training improves performance [7] [32]. Subjects were pre-screened on the CIPIC© HRTF database to find the best-fitting HRTF, with subjects rejected if a suitable HRTF was not available.

3.2.3. Timbre

Timbre is the sound quality or sound color that discriminates, for example, one musical instrument from another. Naturally occurring sounds frequently differ in timbre, which is essential for auditory stream segregation. The salient dimensions of timbre are poorly understood [14] [16] [33]. Yet, timbre is important to the experiment, which uses it to distinguish audio streams.

In the experiment, both pitch and lateral location of audio streams are modulated so these cues are not available for segregation. Timbre, then, is the feature of a stream that allows users to segregate streams. Ideally, streams should have timbres that are as perceptually different as possible. However perceptual difference maximization must respect two constraints that are needed to maintain spatial localizability of the stimuli.

1. Streams need sharp attack with well-articulated transients [34].

2. The frequency spectrum of the streams needs significant energy near 7kHz [7].

The stimuli exploit four specific timbral dimensions, the two first of which are acknowledged being among the most recognizable for audition [31]: brightness of the spectrum, bite of the attack, temporal staticity of the spectrum, and roughness of the sound [18] [33].

The timbres used have different and unique traits on these dimensions. However, even with this latitude it is impossible to find orthogonal timbres that possess adequate segregation. So, the ultimate test for validating segregation among streams is empirical, whether or not typical users can segregate them. Much effort went into equalizing and compressing the dynamics of each timbre to optimize their mutual distinctiveness and to make them pleasant listening. The resulting timbres are based on three natural instruments: the tenor saxophone, the acoustic guitar and the orchestral strings. The spectrograms and temporal envelopes of these timbres are shown in Figure 2.

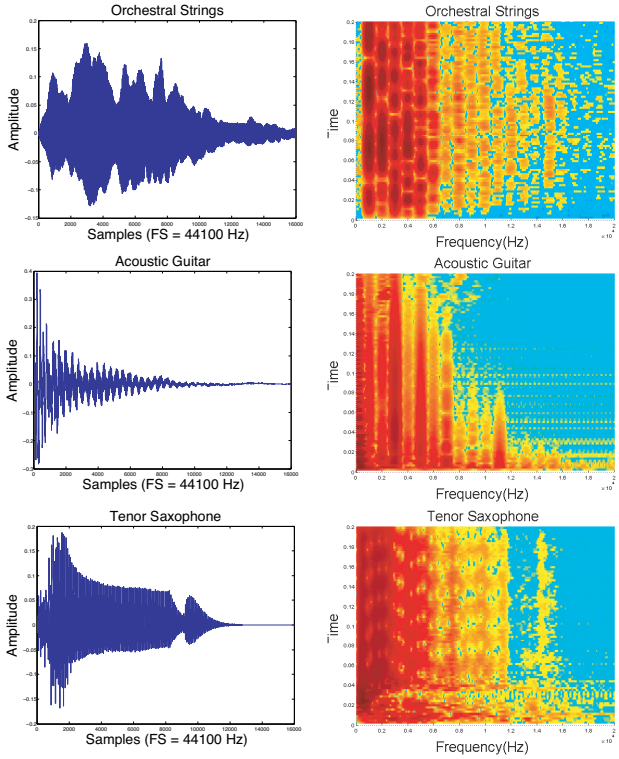

Figure 2: Each row correspond a timbre. From top to bottom: Orchestral Strings, Acoustical Guitar and Tenor Saxophone. The left column shows the envelopes, and the right column the spectrograms.

3.2.4. Velocity

Failure to detect motion is inversely related to source velocity and integration time [21], the minimum integration time (MIT) being the shortest time duration a moving audio stream must persist for motion to be perceived. Fine velocity discriminations are made by successive comparisons of an object's position, at a rates between 1.67 Hz and 6.67 Hz [21]. Consequently, the dominant bottleneck is the rate at which auditory snapshots can be taken [22]. During the experiment, the information delivery rate - the speed at which the interfaces refresh the information - is held constant at 4 Hz. Consequently, new information is displayed every 250ms. This pace is chosen to accommodate MIT. The angular distance traversed per quarter second for each stream when trajectories have an average velocity of 30.3°/*s*, about four displacements of 7.6° each second. The displacements are larger than the MAMA for this velocity [22].

Trajectories having a single stream are unison melodies; trajectories with two streams are musical intervals, two sounds played simultaneously; and trajectories with three streams are triads, three note chords. Depending on the encoding method, some streams are modulated in pitch.

4. EXPERIMENT

In this section we discuss how the participants were selected and how the experiment was conducted. Then, we discuss the audio files' distribution, the environment, the equipment, how data were collected. Finally we justify our choice of trajectories.

4.1. Subjects

Subjects were selected following a pre-test in which the best-fit HRTF was found for every potential subject. Subjects were rejected if it proved impossible to find an HRTF in the CIPIC© database that provided adequate spatial localization. The pretest is essential because best-fit HRTFs work poorly for some subjects. This compromise is not a limitation on the practical utility of the encodings because ongoing users would surely obtain individual HRTFs.

Trajectories like the ones shown in Figure 3 were built to test a subject's ability to discriminate azimuth and elevation. Roughly, the pre-test consisted of:

1. showing a silent QuickTime™ movie showing a trajectory,
2. playing in turn 42 audio versions of the trajectory without visual information, each convolved with one of the 42 HRTF individual captures provided by CIPIC©, and
3. asking the subject which trajectory best portrayed the trajectory they were shown.

The participants could watch again the soundless QuickTime™ movies on demand, and listen more than once any HRTF individual capture of the trajectory. Most subjects identified more than one candidate HRTF individual capture during the pre-test first round, in that case we proceeded by elimination towards the best choices by using another trajectory test, and so on, until the best HRTF measurement was found, if available.

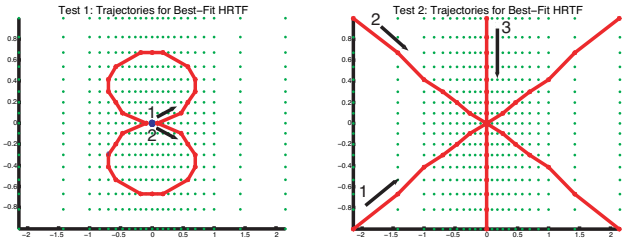

Figure 3: These figures show two trajectories used for screening the participants. Left: the upper circle was played in the direction pointed by *arrow 1* (top), then after the lower circle would be played following the direction pointed by *arrow 2* (bottom), each trajectory separated by 0.5 *s* of silence and both starting and ending at the centre. Right: similar to left, but linear trajectories instead.

Ten paid subjects ran the pre-test but only seven paid Computer Science students, six males and one female, all between 21-26 years old, were selected for the final experiment. Aware that significant performance difference between listeners with and without musical training, not to mention tone-deaf individuals may exist [35], we tested the selected group against tone-deaf deficiency. We report that six of the participants had musical training and none of them were tone-deaf.

4.2. Experiment Procedure

The experiment was performed in a single one hour session, consisting of four sessions, one session per method. Each session began with a brief introduction explaining the method, followed by six test trials. Provided as part of the test trials were simple visual animations showing the path of the test sound, allowing the subjects to see what they were hearing. The test trials were self-paced: participants were allowed to redo the trials in any order, as many times they wanted.

The test trials were followed by twenty-four experiment trials. For each trial, the subject listened to a five second audio file, and responded by giving the trajectories heard in the file. Subjects could play each file as many times they wished, a count of the number of plays being kept. No solutions were provided for experiment trials. All participants completed the experiment in less than one hour. The sessions were presented in a consistent order: PE, BF.HRTF, BF.HRTF&PE, and PE&BF.HRTFix, each of which is explained below.

4.2.1. Audio files

Single stream trajectories were presented on 50% of the experiment trials; two stream trajectories were 33.3% and three stream trajectories were the remaining 16.7%. These trials were randomly intermixed, so subjects did not know how many streams were actually present on any trial. Subjects were instructed to focus on the trajectory as a whole, reporting the shape(s) drawn.

4.2.2. Environment

The experiment was framed in a webpage, presented on the screen of a standard Pentium 4 desktop computer with Firefox™ and QuickTime™. The audio was transmitted from the computer's audio card, calibrated flat, through a pair of SONY professional MDR-7506™ headphones. The experiment was conducted in a quiet environment, one participant at a time.

4.2.3. Responses

At the end of each experiment trial, the subject was presented with seven possible answers, and asked to choose the one that best corresponded to what they heard. One of the answers was the correct solution, the other six were picked using a random weighted sampling method. Each trajectory source/target was weighted inversely proportional to its disparity (see section 5.1 for more), which can intuitively be understood as probability, e.g., the weight for hearing letter 'B', given that 'A' is played would be $P(\text{'B'}|\text{'A'})$. This procedure ensures that trajectories similar to the one played dominate the wrong answers, which increases the discrimination of the data analysis.

4.2.4. Trajectories

All trajectories consist of stroke-based letters and digits. This experiment measures the ability of listeners to detect the shape defined by a motion, without concern for its absolute position or

size. Thus, the highly overlearned human ability to recognize letter forms regardless of translation and scale makes them ideally suited for this experiment.

4.3. Method 1: PE

The first method encodes azimuth by stereo panning and elevation using the Pratt effect (PE), which is based on a linguistic analogy. Using 'high' to describe the sound of short wavelength audio waves, and 'low' for the sound of long wavelength ones is a feature of most languages, which suggests an inherent association between pitch and spatial elevation [32] [36] [37]. In the early nineteen thirties [36] Pratt showed that subjects could localize five pure tones, at octave intervals between 256 Hz and 4096 Hz at elevations from low to high. Ferguson and Cabrera [37] demonstrated the same effect using music stimuli (violin), with some subjects commenting that audio image elevation changed as the melody moved. Unfortunately, ten to twenty percent of the population doesn't hear PE [4]. To put this in context, however, twenty-five percent of the population perceives elevation poorly with HRTF, because of the form of their outer ear [38]. This individual variation may explain some of the benefits of the redundant coding in Methods 3 and 4.

PE is implemented for elevation by associating a half-tone increase in pitch between audio pixels that are adjacent in elevation. The seventeen half-tone pitch range goes from G3 to B4, a mid-scaling range that feels comfortable to most humans. Stereo panning presents equal loudness to both ears at the centreline of the 23 pixel grid. At the edges of the grid the differential loudness reaches its maximum. Creation of the sound image is implemented using a pre-computed lookup table that triggers a pre-transposed and pre-panned audio stream for a given coordinate (x, y).

Table 1: Time Domain Model for the PE experiment. Given the (x, y) coordinates we compute elevation and panning using pitch and ILD respectively.

ELEVATION MODEL	PANNING MODEL
G3 is a $tone$ vector @ $\sim$ 196 Hz	$tone_y$ vector for elevation y
$y = [0, 16]$	$x = [0, 22]$
$tone_y$ = **G3** $\left(2^{(\frac{y}{12})}\right)$	$tone_{(y,L)}$ = $\frac{(22-x)}{22}(tone_y)$
	$tone_{(y,R)}$ = $\frac{x}{22}(tone_y)$

4.4. Method 2: BF.HRTF

This method encodes elevation and azimuth together using the best-fitting HRTF from the CIPIC© HRTF database. The database contains the Head Related Impulse Response filters, the time domain equivalents of frequency domain HRTF filters, of 42 subjects. For each subject the appropriate filter is discovered in the pre-test.

BF.HRTF is implemented using a fast lookup table to convert from planar to angular coordinates because angles in interaural-polar coordinates are used to index a subject's HRIR measurements for a specific location. When an audio pixel is activated, the monophonic audio stream tone is filtered by the right and left ear HRIR filters, h_R and h_L. Because h_R and h_L are time domain vectors, convolving the filters with the tone vector alters the frequency spectrum so as to position the audio stream in space.

4.5. Method 3: BF.HRTF&PE

PE and BF.HRTF have well-recognized problems. For example, both degradation in elevation localization and front-back / up-down reversal is common using BF.HRTF. Alternatively, in some circumstances, PE segregates poorly [39]. Possibly redundant coding of elevation, using both BF.HRTF and PE provides a listener with useful consistency information. Thus, this method combines the two. To implement the combination the monophonic audio stream is first altered in pitch, as for PE. The result is convolved with filters h_L and h_R, as for BF.HRTF. Thus, objects high in elevation are represented by high pitch, and their sound appears to the listener to be above the head, because of the HRIR filtering. In addition, transversal trajectories (left/right) should be perceived naturally as they are handled by binaural HRTF.

4.6. Method 4: PE&BF.HRTFix

BF.HRTF&PE is expected to reduce elevation uncertainty, but does not necessarily address the segregation problems of PE. This method, which to our knowledge is novel, is based on a simple observation. Two audio streams having similar pitch and timbre can be segregated better when played in different ears, than when played together in mono [18]. For example, to improve the 'clarity' of their mix, recording sound engineers normally avoid mixing the high-hat and the shakers together, preferring to distribute one on the left channel, the other on the right channel. Since left/right localization is conveyed through panning, and elevation through pitch, possibly HRTF elevation can be used to 'mix' simultaneous audio streams at different elevations, thus improving segregation just as recording engineers do with left/right panning? HRTF elevation is then held constant for each stream. This approach then helps to separate audio streams by attributing to each a unique and constant elevation over the median plane, with PE representing each object's vertical position. With this scheme, lateral position of objects are conveyed by sound localization obtained from binaural HRTF. Figure 4 shows how convoluting HRTF filters with audio streams could enforce segregation, while physical elevation is modelled using Pratt's effect.

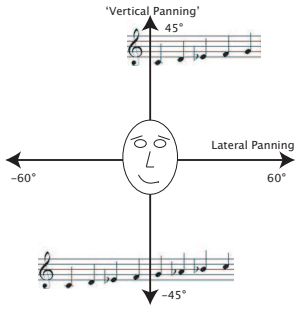

Figure 4: To promote better segregation, the audio streams corresponding to moving objects have their HRTF elevation parameter set constant at distinct elevation. The figure shows two segments corresponding to a scene's objects in motion, each assigned to distinct but constant elevations in 2D space, say -40° for one object and 40° for the other. The physical elevation of objects is translated by pitch inflections represented by the musical scales.

PE&BF.HRTFix is implemented like method 3, with for distinction the elevation which is set constant for each stream, with a unique pre-defined elevation per stream. When an audio pixel is activated, the monophonic audio stream tone is modulated in pitch correspondingly, then filtered by the right and left ear HRIR filters, h_R and h_L at constant elevation value.

5. RESULTS

The raw data is analysed in a two step process: first a closeness score is calculated, which takes into account close matches existing in the subjects' choices; then the score is analysed using conventional analysis of variance.

5.1. Score Calculation

Most experiments in perception score only the fraction correct in subjects' responses. However, because this experiment uses letters and digits as stimuli there is additional information in the wrong answers. For example, if the given trajectory is 'E', an answer of 'F' indicates a closer perception than an answer of 'W'. This information is used to calculate a closeness score for each trial.

Scores are calculated using an Iterative Closest Point (ICP) algorithm [40]. ICP is an iterative descent algorithm that works to minimize the sum of the squared distances between all points in a source trajectory and their closest points in a target trajectory. ICP finds the optimal the translation that best aligns source and target trajectories, then calculates the distance between them. Translation is the only transformation allowed, in order to avoid close matches between, for example, '6' and '9' or '3' and 'E'. ICP deals poorly with outliers, but they are unimportant because of the multiple-choice format of our experiment.

Applying ICP to all trajectories of the possible answers created a confusion matrix, which encapsulates the perceptual distance between all possible pairs of trajectories. A high confusability between two trajectories means that a subject is likely to choose one when the other has been incompletely perceived. Each entry in the confusion matrix is proportional to the inverse of the distance between the two trajectories, normalized so that the sum of each row and column is unity. The entry is, in effect, the score or the probability that the subject will respond with one trajectory given that they incompletely perceived the other.

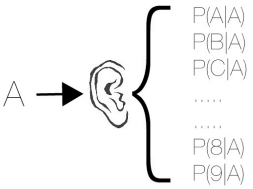

Figure 5: The figure illustrates the score or probability of responding with any trajectory given a specific one, 'A' .

5.2. Speed-Accuracy Trade-Off

In perception, when subjects take longer to judge, their answers are more likely to be correct. This effect is known as the speed-accuracy trade-off. In the experiment subjects were allowed to replay the stimulus as many times as they wished. More replays are expected to be associated with more difficult trials, with subjects taking longer to achieve their desired level of certainty. Thus, the number of replays is expected to be positively correlated with trial difficulty, just as the score is expected to be inversely correlated.

5.3. Analysis of Variance

The closeness score and the number of replays were subjected to analysis of variance, using the same simple procedure in each case. To start an analysis of variance with subject, method and number of streams as factors, and including two-way interactions, was performed. In both cases no interactions were significant and, because subject variation was unimportant, a second analysis using method and number of streams, with no interactions, was performed. The results described are derived from the second analysis.

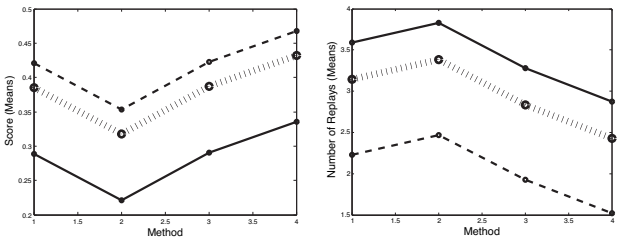

Figure 6: Left shows the average closeness score for different methods. Right: shows the average number of replays for different methods. Read text for details.

Figure 6-left shows the average closeness score for different methods. For all methods the upper point is for one stream, the middle point for two streams and the lowest point for three streams. The difference between methods is marginally significant, ($F(3,162) = 2.42$, $p = 0.068$). The best method is PE&BF.HRTFix; PE and BF.HRTF&PE are approximately equal; BF.HRTF is the worst. Least significant difference post-hoc comparisons show these differences to be marginally significant.

Figure 6-left shows that more streams introduce more error into the perception. The difference between the number of streams is highly significant, ($F(2,162) = 4.75$, $p < 0.01$). The third stream has a larger effect than the second. Post show the difference between two streams and three streams to be significant, ($p = 0.035$).

Figure 6-right shows the average number of replays for different methods. For all methods the lowest point is for one stream, the middle point for two and the upper point for three. The difference between methods is significant, ($F(3,162) = 3.27$, $p = 0.023$). Remembering that lower replays indicate less difficult perceptions the order of methods is the same, with the exception that BF.HRTF&PE outperforms PE.

The number of replays also shows that difficulty increases as the number of streams increases, ($F(2,162) = 11.67$, $p < 0.01$). The agreement in all respects in the results of the two measures increases confidence that the results are robust.

6. CONCLUSION AND DISCUSSION

This paper examined four different methods for the trajectory of an object in an audio signal, two of which are new in this paper. The methods were tested on a group of seven subjects, using two measures of performance. The results show that the fourth method, PE&BF.HRTFix, is significantly superior to the other three. Informal comments of the subjects following the experiment point to the same conclusion.

PE&BF.HRTFix differs from BF.HRTF&PE in the encoding of elevation. The two methods have in common using pitch for elevation and panning for azimuth localization. But PE&BF.HRTFix has the additional feature of using HRTF elevation to separate streams in elevation. This feature helped the participants better to segregate the audio streams, which contributed to making PE&BF.-HRTFix the best solution.

Both measures gave the same ranking for the four methods, suggesting that the results are robust against subjects having different judgement criteria. The measured ranking, from best to worst, is: **PE&BF.HRTFix**, **BF.HRTF&PE**, **PE**, **BF.HRTF** with the caveat that **BR.HRTF&PE** is not statistically different from **PE** .

Probably the biggest surprise in this result is the poor performance of BF.HRTF. Most likely the low-cost pre-test was not effective in getting the best fit for each participant. If so, either anthropometric measurement [41] or individually measured HRTFs would improve performance. Note, however, that these improvements are costly, and the low-cost pre-test was sufficient to give better results with PE&BF.HRTFix. Because most initial use of audio motion detection is casual, with investment deferred until the system has proved its utility, the low investment pre-test is an attractive feature of PE&BF.HRTFix.

A weakness of the experiment design is the confound with possible learning because the task was novel to the subjects. However, it uses perceptual mechanisms that are highly practiced as a result of following moving sounds throughout the subject's prior life. Thus, the basic mechanisms supporting the experimental task are unlikely to be subject to learning during the experiment.

All the same, the subjects may be gaining experience with the exact perception needed for the experiment. This, however, is very unlikely. The session sequence is PE, BF.HRTF, BF.HRTF&-PE, and PE&BF.HRTFix. If learning were to dominate the results, performance would be ordered in this sequence, but this differs from the order of performance, which is BF.HRTF, PE equal BF.HRTF&- PE, and PE&BF.HRTFix best. Thus, learning is not an important effect.

The limited number of subjects from a homogenous population in test is reasonable since we are asserting basic auditory qualities having low variance among humans with normal hearing. We think that musical training helped novice users to discriminate timbres and pitch modulated streams. A user with non-musical training would develop similar skills after extensive use of the interface. The natural step is now to involve blind subjects in future developments.

In summary, this paper introduced two novel methods for encoding motion in an auditory signal, combining the Pratt Effect with a best fit HRTF, and separating the Pratt Effect from the elevation component of the HRTF to improve stream segregation. They were compared experimentally with the Pratt effect alone and with the best fit HRTF alone, where the subject's task was to determine the global motion of a moving source of sound. Both novel methods outperformed best-fit HRTF, an usual method of encoding spatial information in sound without the overhead of individual HRTF measurement.

The experiment examined the perceptually demanding task of determining the motion of simultaneously presented streams encoding different motions, for which both stream segregation and motion following must be performed together. The best performing of the novel methods, PE&BF.HRTFix, used an unusual method for enhancing segregation, separating the HRTF elevation and azimuth information.

This result has more general significance. The first novel encoding, BF.HRTF&PE, uses the well-known principle of redundancy, presenting hard to perceive information in two separate channels which should agree with one another. The benefit is small, suggesting either that subjects have little trouble acquiring adequate elevation information from the Pratt effect alone, or that there is no easy way to combine information from the Pratt effect with elevation information from HRTF. The better peformance of the second novel method, PE&BF.HRTFix, suggests that stream segregation is a bottleneck, which is not surprising given the limited processing capacity of short term memory.

However, there remains a mystery. If stream segregation is the whole story we would expect there to be an interaction in the analysis of variance between the number of streams and the encoding method. PE&BF.HRTFix should show a larger improvement when there are more streams to segregate. Understanding this anomaly is the most important future work suggested by the results reported in this paper. It is probably the key to improving human performance at deducing motion from multiple input streams. The presence of multiple streams is the normal case when vision deprived users access a normal environment through their auditory sense. It also promises deeper theoretical insights into the nature of coding for display. When is it better to use redundant coding to emphasize important attributes of the stimulus? When in the holistic nature of perception better assisted by adding extra, possibly less important, information, even when is comes at the cost of naturalism?

Regardless of the outcome of these theoretical issues, the research introduces a novel method for encoding motion in auditory stimuli, a method that outperforms its competitors on two dimensions, difficulty and correctness. Thus, even in its present form it is a practical solution for improving motion perception in auditory displays. It is clear, however, that improvements are possible, and that making these improvements is an important direction for future research. For example, several restrictions were imposed in order to make the experiment tractable: depth was suppressed, velocities were compressed, the range of elevation was limited, the number of streams was small, and so on. The initial assumption is that the restrictions are arbitrary, and that the method fails on in extreme cases. But 'extreme' needs an operational definition, which can only be created by more extensive experimentation which covers a broader range of conditions. Our future research will follow examine such of these questions, as they arise while we are engineering and testing a fully implemented system and testing it on vision deprived users in a variety of natural settings.

7. ACKNOWLEDGMENTS

The authors would like to thank all the participants for their motivation and commitment. Martin Talbot would like to thank Richard Mann, Pascal Poupart, Claude-Guy Quimper, Heng Yu, Elodie Fourquet, Laurent Charlin and Caroline Charest for taking their time discussing the project and for their interest in its success. Bill Cowan is partially supported by NSERC Discovery Grant and Martin Talbot by NSERC PGS M.

8. REFERENCES

[1] http://www.seeingwithsound.com/

[2] Boff K. R., and Lincoln, J. E. (1988). *Engineering Data Compendium, Human Perception Performance, Volume 1 & 2*, Ohio, USA: Harry G. Armstrong Aerospace Medical Research Laboratory Wright-Patterson Air Force Base.

[3] Kapralos, B., Jenkin, M. R. M. , Milios, E. (2003) "Auditory Perception And Spatial (3D) Auditory Systems," *Technical report CS-2003-07*, York University.

[4] Terasawa, H., Slaney M., Berger J. (2005). "Perceptual Distance In Timbre Space," *Proc. of ICAD 05*, Limerick, Ireland.

[5] Rayleigh, L. (1907). "On Our Perception Of Sound Direction," *Philosophical Magazine*, 13, 214-232.

[6] Batteau, D. W. (1968). *Listening With The Naked Ear*, In S. J. Freedman (Ed.), *The Neurophysiology Of Spatially Oriented Behavior. Homewood*, IL: Dorsey Press.

[7] Begault, D. R. (1994). *3D Sound: For Virtual Reality And Multimedia*, Cambridge, MA: Academic Press Inc.

[8] Jones, D. L., Stanney, K. M., and Foaud, H. (2005). "An Optimized Spatial Audio System for Virtual Training Simulations: Design and Evaluation," *Proc. of ICAD 05*, Limerick, Ireland.

[9] Huopaniemi, J. (1999). "Virtual Acoustics and 3D Sound In Multimedia Signal Processing," *PhD thesis*, Department of Electrical and Communications Engineering, Helsinki University of Technology, Espoo, Finland.

[10] Afonso, A., Katz, B. FG., Blum, A., Jacquemin, C., and Denis, M. (2005). "A Study Of Spatial Cognition In An Immersive Virtual Audio Environment: Comparing Blind And Blindfolded Individuals," *Proc. of ICAD 05*, Limerick, Ireland.

[11] Martens, W. L. (2003). "Perceptual Evaluation of Filters Controlling Source Direction: Customized and Generalized HRTFs for Binaural Synthesis," *Acoust. Sci. and Tech.* 24, 5, 220-232.

[12] http://www.afb.org/afbpress/pub.asp?DocID=AW040204&Mode=Print

[13] http://www.batforblind.co.nz/history.htm

[14] Terasawa, H., Slaney, M., Berger, J. (2005). "Perceptual Distance In Timbre Space," *Proc. of ICAD 05*, Limerick, Ireland.

[15] Grey, J.M., and Moorer, J.A. (1977). "Perceptual Evaluation Of Synthesized Musical Instrument Tones," *J. Acoust. Soc. Am., 62*, 454-462.

[16] Peeters, G., McAdams, S., and Stravinsky, I. (2000). "Instrument Sound Description In The Context Of MPEG-7," *Proc. of ICMC2000 (International Computer Music Conference)*, Berlin, Germany.

[17] Deutsch, D. (1972). "Octave Generalization And Tune Recognition." *Perception & Psychophysics*, 11, 411-412.

[18] Bregman A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*, Cambridge, MA: MIT press.

[19] Jones, J. R. (1976). "Time, Our Lost Dimension: Toward A New Theory Of Perception, Attention And Memory." *Psychological Review*, 83, 323-355.

[20] Brant, H. (1967). *Space As An Essential Aspect Of Music Composition. Contemporary Composers on Contemporary Music*, NY: Holt, Rinehart & Winston.

[21] Grantham, D. W. (1986). "Detection and Discrimination of Simulated Motion of Auditory Targets in the Horizontal Plane." *J. Acous. Soc. Am.* 79, 1939-1949.

[22] Chandler, D. W., Grantham, D. W. (1992). "Minimum Audible Movement Angle In The Horizontal Plane As A Function Of Stimulus Frequency And Bandwidth, Source Azimuth, And Velocity," *J. Acout. Soc. Am.* 91, 3, 1624-36

[23] Strybel, T. Z., Witty, A. M., and Perrott, D. R. (1992). "Auditory Apparent Motion in the Free Field: The Effects of Stimulus Duration and Intensity." *Percep. & Psycho.* 52, 2, 139-143.

[24] Valjamae, A., Larsson, P., Vastfjall, D., and Kleiner, M. (2005). "Travelling Without Moving: Auditory Scene Cues For Translational Self-Motion," *Proc. of ICAD 05*.

[25] Michotte, A. (1963). *The Perception of Casualty*, New York: Basic Books Inc.

[26] O'Leary, A., and Rhodes, G. (1984). "Cross-Modal Effects on Visual and Auditory Object Perception." *Percep. and Psycho.* 35, 565-569.

[27] Susini, P., McAdams, S., Smith, B. K. (2005). "Loudness Asymmetries For Tones Increasing And Decreasing Levels," *Proceedings of ICAD 05*, Limerick, Ireland.

[28] Hartmann, W. M. (1974). "Localization of Sound in Rooms," *J. Acout. Soc. Am.*, 74, 1380-1834.

[29] Su, T. K., and Recanzone, G. H. (2001) "Differential Effect Of Near-Threshold Stimulus Intensities On Sound Localization Performance In Azimuth And Elevation In Normal Human Subjects," *J. of the Association for Research in Otolaryngology*, 2, 3, 246-256.

[30] http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm

[31] Strybel, T. Z., Manligas, C. L., Perrott, D. R. (1992). "Minimum Audible Movement Angle as a Function of the Azimuth and Elevation of the Source," *Human Factors*, 34, 3, 267-275.

[32] Scruton, R. (1983). "Understanding Music," *Ratio*, 25, 2, 97-120.

[33] Ricard, J. (2004) "Towards Computational Morphological Description Of Sound," *PhD. dissertation*, Univesitat Popeu Fabra.

[34] Hartmann, M. W. (1983). "Localization Of Sound In Rooms," *J. Acout. Soc. Am.* 74, 1380 -1391.

[35] Cabrera, D., Tilley, S. (2003). "Parameters For Auditory Display Of Height And Size," *Proc. ICAD 2003*, Boston, USA.

[36] Roffler S. K., and Butler, R. A., (1968). "Localization Of Tonal Stimuli In The Vertical Plane," *J. Acoust. Soc. Am.*, 43, 6, 1260-1266.

[37] Ferguson S., and Cabrera, D. (2005). "Vertical Localization Of Sound From Multi-Way Loudspeakers," *J. Audio Eng. Soc.*, 53, 163-173.

[38] Noble, W. (1987). "Auditory Localization In The Vertical Plane: Accuracy And Constraint On Bodily Movement," *J. Acoust. Soc. Am.*, 82, 5, 1631-1636.

[39] Best, V. et al. (2003). "Spatial Effect Of The Segregation For Sounds In Virtual Auditory Space," *Proc. 8th Western Pacific Acoustics Conf*, Melbourne, Australia.

[40] Besl, P. J., and McKay, N. D. (1992). "A Method For Registration Of 3-D Shapes," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* 14, 2, 239-256.

[41] Algazi, V. R., Duda, R. O., Thompson D. M, and Avendano, C. (2001). "The Cipic Hrtf Database," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.