Mapping the perceptual topology of auditory space permits the creation of hyperstable virtual acoustic environments

W. Owen Brimijoin*, Shawn Featherly† and Philip Robinson‡
Facebook Reality Labs, Redmond, WA, USA

Abstract: The perception of acoustic motion is not uniform as a function of azimuth; listeners need roughly twice as much motion at the side than at the front to judge the two motions as equivalent. Self-generated acoustic motion perception has also been shown to be distorted. Sounds moved slightly with the listener’s head are more consistently judged to be world-stable than those that are truly static. These distortions can be captured by a model that incorporates a head-centric warping of perceived sound location, characterized by a displacement in apparent sound location away from the acoustic midline. Such a distortion has been demonstrated; listeners tend to overestimate azimuth when they are asked to point at a sound source while keeping their head and eyes fixated ahead of them. Here we show that this mathematical framework may be inverted and we demonstrate the benefits of re-mapping sound source locations toward the auditory midline. We show that listeners prefer different amounts of spatial remapping, but none preferred no remapping. Modelling shows minimal impact on spatial release from masking for small amounts of remapping, demonstrating that it is possible to achieve a more stable perceptual environment without sacrificing speech intelligibility in spatially complex environments.

Keywords: Sound localization, Spatial hearing, Self motion, Motion perception, Hyperstability

PACS number: 43.66.Pn, 43.66.Qp  [doi:10.1250/ast.41.245]

1. INTRODUCTION

The acoustic world is not as spatially stable as it appears to be. This is true for real-world signals [1,2], but sound sources appear to be particularly unstable in virtual acoustic environments, especially in anechoic conditions. This is most apparent when turning the head and noting the degree to which a signal at the front appears to move with respect to the head. Many listeners will judge a sound at front to be moved too much, as if the virtual environment were being counterrotated against the head turn by too large a degree.

The reasons for this are unclear but appear to be related to a systematic distortion in our perception of acoustic space. Listeners make significant errors when asked to identify the direction of arrival of a sound or to make judgements about the movement of a sound source. These errors tend to increase as the sound source is moved away from the acoustic midline [3]. Critically, however, the errors consist not only of scatter around a central point, they also include systematic offsets or perceptual biases. These offsets form the basis of a perceptual topology of acoustic space, a term likely first used by Oldfield and Parker [4].

It is known that this topology is warped by a number of factors, including eye position [5] and head-to-trunk angle [6]. One example of a systematic bias is that listeners routinely overestimate the direction of arrival of static signals [4,7–9]. That is, listeners tend to judge sound sources as being at a greater angle away from their acoustic midline than they truly are. Recently published results suggest that this static distortion is related to strong changes in a listener’s perception of auditory motion [10], consisting of an expansion space at the front of the listener and a compression at the side.

Mathematically quantified, this relative expansion and contraction suggests that the perceptual location of static sound sources is warped with respect to the head. Since the observed distortion can be described mathematically, this gives us a straightforward means by which we can compensate for it. An inversion of the function was published previously [10]; here it is presented as Eq. (1).

\[\theta_a = \frac{90 \times \ln(10) \times \tanh^{-1}\left(\theta_p \tanh\left(\frac{\ln(Rt - c)}{\ln(10)}\right) / 90\right)}{\ln(Rt - c)}\]

---

*e-mail: owen.brimijoin@oculus.com
†e-mail: shawn.featherly@oculus.com
‡e-mail: philrob22@oculus.com
In this function, all angles are in degrees, tanh is computed for radians, ln is the natural logarithm, $r$ is a constant equal to 7.08, $c$ is a constant equal to 5.97, $\theta_a$ is the actual position of a signal, and $\theta_p$ is the perceived position of that signal. $R$ here is the ratio between the point of subjective equality for angular excursion at 90° and 0°. This may be thought of as an ‘expansion ratio, specifying the amount of spatial expansion as a function of azimuth. Increasing $R$ results in signal source angles being moved closer to the listener’s midline, away from the sides, thus compressing space at the front and thereby decreasing the rate at which signals at the front appear to move as the head moves.

The goal in this study was to determine what preferences listeners might have for the expansion constant $R$ in the equation. This required a pilot experiment to allow listeners to fine tune the $R$ value until they were satisfied that it resulted in the most perceptually stable virtual environment. Any implementation would be simplest if it did not require any individualization, so we sought to find a single $R$ value that would be likely to be accepted by the largest number of listeners as an improvement over standard geometric presentation.

2. METHODS

Listeners ($N = 60$, all employees of FRL, ages were not recorded) were seated in a quiet room and given an Oculus Rift head-mounted display to wear. They were presented with a virtual room with a wooden console on top of which was placed a radio. A static depiction of the virtual environment is shown in Fig. 1.

The radio emitted a sound that consisted of a quiet radio program mixed with a sawtooth-modulated pink noise (50% amplitude modulation at 6 Hz), spatialized using the Unity engine and presented over Oculus Rift headphones. This signal was sufficiently broadband and contained enough onset transients to be easily localizable.

All sounds were presented at a user-determined presentation level (not recorded). The signals were presented anechoically, with no attempt to simulate the room acoustics of the listener’s perceived environment.

They were allowed use the Touch controller trigger button to toggle back and forth between (A) standard Rift SDK audio (i.e., geometrically correct rendering) and (B) a spatially warped version. Both the virtual environment and spatial warping were implemented in Unity (in C#). Participants controlled the slider in the B version with the thumbstick on the Touch controller, which in turn adjusted the value of $R$ in steps of 0.1, with the far left $R$ value matching standard Rift geometry.

The listeners were asked to move the slider back and forth until they felt that the audio location did not move in either direction relative to the radio when they turned their heads. They were allowed at any time to compare the warped version with standard Rift geometry presentation. They indicated that they were satisfied that the slider was in the ideal position for them by pressing A and B together on the Touch controller. While this method of self-adjustment is not the most psychophysically robust means of measuring a point on a psychometric function, it has the benefit of being easily and quickly measured, making the relatively large $N$ possible in a short pilot study.

After recording the listener’s preferred $R$ value, the participants were then asked to judge how sure they were of their assessment. Specifically, they were asked how confident they were that they had chosen the ideal point on the slider for them. The confidence assessment was presented as a Likert scale with options ‘Not at all confident,’ ‘Slightly confident,’ ‘Moderately confident,’ ‘Very confident,’ and ‘Completely confident.’ These were mapped to integer values from 0–4. These data were analyzed as follows. The selected ratio values were binned in increments of 0.5 and plotted as a histogram. For the confidence values, the mean was taken for confidence values in each 0.5 Ratio bin, excluding bins in which there were fewer than 3 measured confidence values.

3. RESULTS

Listeners were variable in the time it took them to choose a value, but took on average 170 seconds (STD ± 98 s). Figure 2 shows the histogram of the user-selected Ratio values. It can be seen that listeners were most likely to choose Ratio values between 3.0 and 4.5. No listeners chose a value equal to 1.0 (standard Rift geometry with no expansion). These are relatively large values of expansion as compared to predictions from published data which suggest an ideal value closer to 2 (Brimijoin, 2018). It was critical, therefore that we assessed both the confidence with which listeners made their judgement and the potential impact of large Ratio values on speech.
intelligibility (see speech intelligibility discussion below). The mean listener confidence values are superimposed on top of the histogram (orange line and right hand y-axis).

Confidence values were highest for listeners that chose Ratios between 2.0 and 3.0, reaching a maximum value greater than 3, corresponding to a confidence label of ‘Very Confident.’ Listeners who felt most strongly about their choice chose $R$ values that were smaller than the group average. Correspondingly, those listeners who chose values near the mean were less confident in their assessment, averaging values between 1 and 2 labelled as ‘Slightly Confident’ or ‘Moderately Confident.’ This suggests that to satisfy most listeners, an $R$ value smaller than the listener mean could be chosen without negatively affecting most listeners’ preference in stabilization.

The spatial warping shifts the location of signal sources closer to the midline. This could have an impact on intelligibility for sources towards the front of the listeners, since any masking noises in the environment would be brought closer to the source of interest. When masker signals are further from the signal of interest, the result is an increase in signal intelligibility that is referred to as spatial release from masking (SRM). Thus the hyper-stability algorithm, by moving the presented angle of an off-axis masker in towards a signal at 0 degrees, would reduce the amount of SRM. In other configurations, such as when a target signal is at 45 degrees, there would be essentially no impact since the equation is relatively flat in this range of angles, and for a target signal out to the side at 90 the expansion correction would actually serve to increase intelligibility for that signal by moving sound-source locations apart from each other. Taking, however, the worst case scenario for a signal of interest at 0 degrees and a noise masker at more eccentric positions, we ran simulations using Bronkhorst’s [11] model of spatial release from masking. The results of these simulations are shown in Fig. 3.
Using an $R$ value of 3 in Fig. 3(A), we can see that as compared to standard presentation methods (green), using a hyperstability correction (orange) results in a slight decrease in SRM. The loss in SRM is plotted in black, showing that for $R = 3$, the loss of SRM peaks at just over 1 dB at a masker location of about 45 degrees. Figure 3(B) shows the SRM loss for $R$ values between 1 and 7. In all cases the maximum SRM loss occurs for targets at 0 and maskers between 40 and 50 degrees. This maximum loss is plotted in Fig. 3(C) as a function of $R$ value. If we seek to minimize the impact of stabilization on intelligibility, it would be best to limit ourselves to having less than a 1 dB maximal decrease in SRM. This occurs at an $R$ value just greater than 2.5.

The choice of a single $R$ value for the hyperstabilization algorithm should be determined by each of these factors, and we sought to satisfy listener preferences, preference confidence, and worst-case scenario speech intelligibility constraints. Given these factors an $R$ value of approximately 2.5 appears to be a good compromise between minimal SRM loss, maximum listener confidence, and mean listener preference.

4. DISCUSSION

Listeners experience distortions in their perception of where sounds are located in space, and how they move with respect to the head and even eye position [5]. This distortion has consequences for the perception of sounds sources by moving listeners, resulting in an apparent mismatch between how sounds move relative to their own motions. For reasons that we do not yet understand, this mismatch is more apparent in virtual acoustics than in the real world. This could be attributed to the relative scarcity of sensory cues in a virtual environment, especially in an anechoic rendering, but future studies may need to address this by varying realism across multi-sensory cues.

In the immediate term, however, we have demonstrated that listeners prefer sound sources that are moved slightly with their heads when in the front, and slightly against their heads when at the sides. This spatial warping results in a more perceptually stable acoustic environment. Future work will address outstanding issues concerning the influence of room reverberation, the presence of multiple sound sources, gaze angle, and apparent distance on ideal $R$ values, but for now we argue that it is possible to use distortion compensation to enable head-tracked audio that more closely matches the location of sound sources and the movement of the visual scene.

REFERENCES