# IN EAR TO OUT THERE: A MAGNITUDE BASED PARAMETERIZATION SCHEME FOR SOUND SOURCE EXTERNALIZATION

*Griffin D. Romigh, Brian D. Simpson, Nandini Iyer*

711th Human Performance Wing
Air Force Research Laboratory
2610 7th street, WPAFB, OH 45344, USA
griffin.romigh@us.af.mil

## ABSTRACT

While several potential auditory cues responsible for sound source externalization have been identified, less work has gone into providing a simple and robust way of manipulating perceived externalization. The current work describes a simple approach for parametrically modifying individualized head-related transfer function spectra that results in a systematic change in the perceived externalization of a sound source. Methods and results from a subjective evaluation validating the technique are presented, and further discussion relates the current method to previously identified cues for auditory distance perception.

## 1. INTRODUCTION

Most spatial auditory displays are built around the understanding that, by replicating the natural cues listeners use to determine a sound source's location, any single-channel sound source can be imbued with spatial attributes. These cues are the complex function of space, frequency, and individual listener known as a Head-Related Transfer Function (HRTF). HRTFs capture the acoustic transformation a sound undergoes as it travels from a specific location in space, interacts with a listener's head, shoulders, and outer ears, and arrives separately at the two eardrums [1]. A single-channel sound filtered with the right-ear and left-ear HRTF corresponding to a specific location can then be presented over headphones with directional accuracy and fidelity comparable to a free-field source provided the HRTF measurements were individualized to the listener [2, 3, 4].

Unfortunately, due to logistical issues associated with acquiring individualized HRTF measurements, most spatial displays utilize non-individualized HRTFs or subsets of the cues contained therein (e.g., only gross binaural cues), resulting in less perceptually accurate spatial representations. A frequent bane for headphone-based displays is the problem of poor sound source externalization. Sometimes referred to as "inside-the-head-locatedness" [5], poor externalization (or internalization) is the perceptual phenomena where sound sources are perceived to originate from a location within a listener's head rather than from out in space where the display designer had intended. Blauert [5] was one of the first to throughly review the literature associated with externalization and suggested that the lack of externalization was merely an endpoint on the continuum of perceived auditory distance (i.e. a sound source appears so close to you that it is perceived inside your head).

The theory that perceived externalization is part of auditory distance perception was backed up by the experimental evidence of Hartmann and Wittenberg [6]. They showed that perceived externalization could be manipulated systematically for a harmonic tone complex by zeroing out inter-aural phase differences (IPD) for tonal components above a given frequency; the lower the critical frequency the less externalized the sound source was judged to be up to a cutoff frequency near 1kHz. Hartmann and Wittenberg [6] also showed that externalization was not affected by forcing a single frequency independent interaural timing difference (ITD) cue, and that both monaural magnitude spectra need to be preserved across the entire frequency range to ensure good externalization not just the gross interaural level difference (ILD). Proper externalization (and/or distance perception) has also been liked to a number of other factors including the use of dynamic head-motion cues [7], ratio of direct to reverberant energy [8], and the high frequency roll off [9].

Despite the previous efforts, it is not clear what the role of spectral features are in perceived externalization. At the extremes, there is a very clear relationship between the presence of monaural spectral features and externalization, such that an absence of spectral features when implementing only a frequency independent ILD causes poor externalization, while the full representation of the spectral features when implementing an individualized HRTF produces good externalization. It is less clear however what is perceived with compressed spectra, where the narrowband spectral cues are present yet potentially diminished in magnitude. Based on that question and the desire for a simple parameterization with which externalization could be effectively modulated, the current investigation aimed at determining whether externalization could be reliably controlled through simple modifications to the monaural magnitude spectrum contained in an HRTF.

## 2. PARAMETERIZATION

From previous literature it is clear that the two extremes of externalization can be attained using methods that differ only in spectral representation of their spatial filters. On one hand, if spatial information is imparted on a sound source using only binaural cues (ILD and ITD), the sound source will appear as though it originates from somewhere along the interaural axis inside the listener's head [10]. On the other hand, well localized *and* externalized sound

DOI: 10.21785/icad2016.035

sources can be created virtually if the source is rendered with an individualized HRTF that preserves the ITD and both left-ear and right-ear monaural spectral cues [11]. These two methods of spatial presentation differ only in the way the spectrum of the sound source at each ear is modified. The parameterization detailed in this section provides a straightforward method to linearly interpolate between the spectra that are known to result in good externalization and spectra which are known to result in strong internalization.

Starting from an individualized HRTF measurement, the following method systematically varies the prominence of spectral features of the left-ear log-power spectrum; an identical procedure is used to modify the right-ear spectrum. If we define $\mathcal{H}^L_{\phi,\theta}[k]$ to be the individualized left-ear log-power spectrum (i.e. decibel scale) corresponding to a location with azimuth $-180^o \leq \phi < 180^o$ and elevation $-90^o \leq \phi < 90^o$, the location-specific, frequency-independent average monaural level $\mathcal{A}^L_{\phi,\theta}$ is defined as in Eq. 1.

$$\mathcal{A}^L_{\phi,\theta} = \frac{1}{K}\sum_k \mathcal{H}^L_{\phi,\theta}[k] \tag{1}$$

Here, $k$ is used to represent one of $K$ discrete frequency values in the valid positive frequency range for the HRTF. For the present work $2 \leq k \leq 174$, making $K = 173$, which represents the positive frequencies from approximately 200 Hz to 15 kHz for a 512 length DFT at a sampling rate of 44.1 kHz. This average level is subtracted from the measured HRTF to give the left spectrum $\mathcal{S}^L_{\phi,\theta}[k]$ as in Eq. 2, which contains all of the spectral features.

$$\mathcal{S}^L_{\phi,\theta}[k] = \mathcal{H}^L_{\phi,\theta}[k] - \mathcal{A}^L_{\phi,\theta} \tag{2}$$

These two components are then weighted and recombined to give the transformed spectrum $\tilde{\mathcal{H}}^L_{\phi,\theta}[k]$ as in Eq. 3

$$\tilde{\mathcal{H}}^L_{\phi,\theta}[k] = \frac{\alpha}{100}\mathcal{S}^L_{\phi,\theta}[k] + \mathcal{A}^L_{\phi,\theta} \tag{3}$$

The parameter $\alpha$ in Eq. 3 varies the magnitude of the spectral features contained in the final transformed spectrum $\tilde{\mathcal{H}}^L_{\phi,\theta}[k]$ from full scale for $\alpha = 100$ to nil for $\alpha = 0$. The transformed spectra corresponding to several levels of the $\alpha$ parameter are shown in Fig. 1 for the left and right ears of a representative subject at two locations on the horizontal plane.

## 3. METHODS

### 3.1. Subjects

Eight paid listeners (5 males, 3 females) with normal audiometric thresholds participated in the subjective evaluation experiment. Listeners participated in 60 trials broken into self-paced 30 minute blocks over the course of two weeks. All listeners had previous experience with virtual spatial audio in the context of objective localization experiments conducted within the laboratory, however, all listeners were believed to be naive to both the subjective evaluation method presented below and to the formalized concept of externalization at the onset of the experiment. As such, the subjects were presented with the following verbal description of externalization at the onset of every experimental trial to familiarize them with the question at hand.

> **Externalization:** *To what extent does the virtual source sound outside your head?*
>
> In this type of trial you will be asked to judge how each virtual source is positioned relative to yourself. When listening over headphones, some sounds may appear as though they originate from inside your head (completely internalized) while others may sound as though they clearly come from a physical location out in space (completely externalized); variations between these two extremes are also possible where a sound might appear to come from on your face, head, or neck or just outside your body.

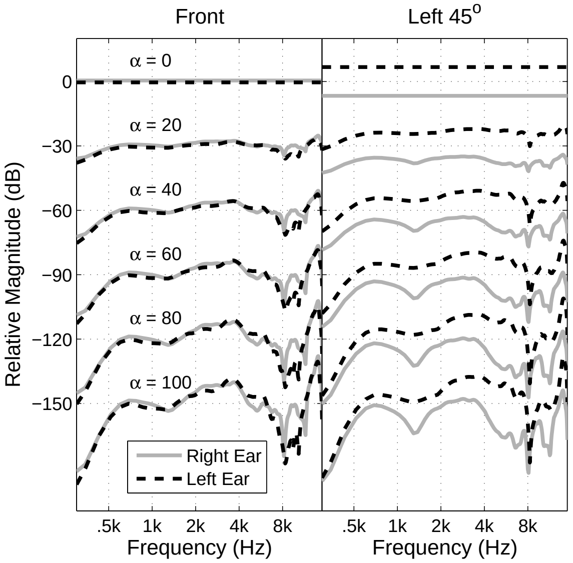


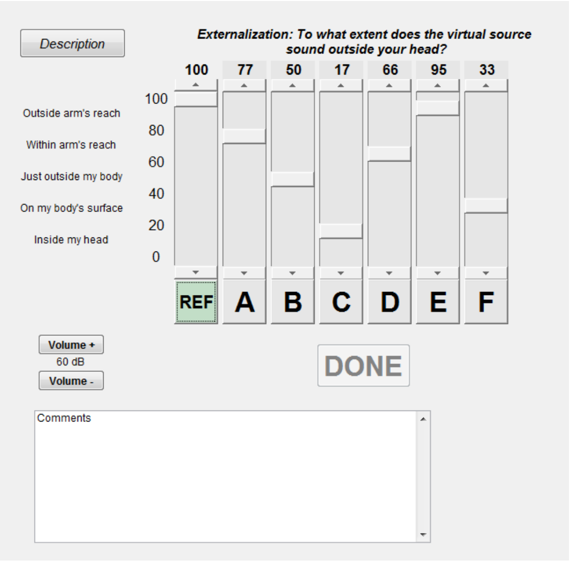
Figure 1: Transformed HRTF spectra for a single subject at to locations on the horizontal plane (panels) as a function of the $\alpha$ parameter. Spectra were given a $\frac{-30\alpha}{20}$ dB gain for plotting purposes.

### 3.2. Task Description

A single experimental trial consisted of an implementation of the Multi-Stimulus Test with Hidden Reference and Anchor (MUSHRA) [12]. In this task, listeners were presented with the GUI depicted in Fig. 2. By pressing the selection buttons on the GUI (labeled "REF","A","B",..."F"), listeners were able to selectively listen to each stimulus one at a time as many times as they desired and in any order throughout a trial. Listeners were asked to compare the various stimuli both to each other and to a reference stimulus and provide an externalization rating for each stimulus according to the scale in Table 1 which was always visible to the subjects at the left of the GUI. They were also were provided written instructions on the use of the GUI and informed that the reference stimulus should correspond to a rating of 100 on the provided scale. At any time during a trial they could reexamine the verbal description of externalization, adjust the overall stimulus level, and leave comments utilizing the GUI.

### 3.3. Stimuli

On a given trial seven different stimuli were employed, the reference stimulus and six test stimuli. The reference stimulus always consisted of a virtual sound source rendered with a full HRTF,



Externalization: To what extent does the virtual source sound outside your head?

Figure 2: Graphical user interface for the MUSHRA task used during the subjective evaluations.

| Rating | Description |
|---|---|
| 80 - 100 | Outside arms reach |
| 60 - 80 | Within arms reach |
| 40 - 60 | Just outside my body |
| 20 - 40 | On the surface of my face, head, or neck |
| 0 - 20 | Inside my head |

Table 1: Verbal descriptions of different levels of externalization and the corresponding rating values used for the subjective evaluation.

while the test stimuli consisted of virtual sound sources rendered with HRTFs that had been transformed as described in Sec. 2 for $\alpha$ values of 0, 20, 40, 60, 80, and 100. The $\alpha = 100$ stimulus was identical to the reference stimulus, therefore acting as the hidden reference and the $\alpha = 0$ stimulus acted as the hidden anchor.

All individualized HRTFs had been previously measured on each listener with the methods described in [3] and consisted of 256 minimum-phase DFT coefficients (sampled at 44.1 kHz) for each ear and a corresponding ITD value at 2 spatial locations, in front of and $45^o$ to the left of the subject). HRTFs were converted to the log-power (decibel) domain, transformed, and ultimately converted back to 256-tap minimum-phase filters. Test stimuli, each 5 s in duration, were generated by convolving one of three single-channel base signals (broadband noise, music, spoken English) with the resulting right and left filters and delaying the contralateral ear by the ITD value. The music and speech samples were taken from the Sound Quality Assessment Material recordings for subjective tests (Track 70 and Track 50, respectively) [13]. All stimuli were normalized post-filtering to have the same average initial level of 60 dB SPL. Each subject participated in ten trials for each location and stimulus type for a total of 60 trials.

## 4. RESULTS

Across all three base stimuli and both azimuths, two subjects showed a negative correlation with the average trend ( r = -0.78 , r = -0.81) computed with their data removed. While is it conceivable that their reference HRTF produced poorly externalized stimuli through some type of measurement artifact (thus resulting in a low externalization rating for $alpha = 100$) , the near perfect reverse ratings exhibited, including rating the anchor, which contained only gross ITD and ILD cues, with a high externalization rating, leads the authors to believe that the subjects misunderstood the task or rating scale. Because of this these two outliers were removed from from remaining data and analysis.

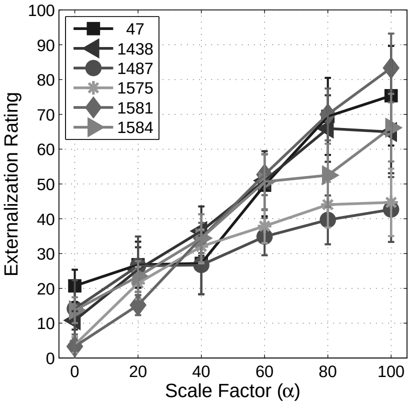


Figure 3: Average externalization ratings for the filtered noise base stimuli at the left $45^o$ position by subject. Error bars represent 95% confidence intervals for the pooled data.

Figure 3 shows the average ratings for the remaining subjects at the left $45^o$ azimuth location averaged across the three base stimuli as a function of the $\alpha$ level. It is clear from Fig. 3 that subjects showed a systematic linear relationship between the $\alpha$ parameter and their externalization ratings. Repeated measures ANOVA results indicates a significant main effect for the $\alpha$ parameter on the externalization rating (F(5,5) = 9.073, $p < 0.001$). In general subjects indicated the $\alpha = 0$ condition (containing only ITD and ILD cues) was “Inside (the) head” or “On the surface”. Less consensus is seen in the slope of the functions however, and likewise the rating given to the full HRTF condition ($\alpha = 100$), where ratings varied from “Outside arms reach” to “Just outside (the) body”.

In contrast, ANOVA results did not indicate a statistically significant main effect for base stimulus type (F(2,2) = 2.305, p = 0.150). Results comparing the ratings for the three base stimuli types averaged across subjects and location are shown in Fig. 4. While not statistically significant, the average data does show a slight trend for externalization ratings to be highest in speech condition compared to the music for bandpass filtered noise.

Figure 5 shows the ratings for the two locations as a function of the $\alpha$ parameter averaged across subjects and stimulus type. Clearly evident in the figure is an interaction with the stimulus location and the $\alpha$ level; the ratings start lower but increase more rapidly for the left 45° location compared to the front. This observation is backed up by ANOVA results which show a significant interaction for $\alpha$ and location (F(5,5) = 4.891, p = 0.003), but

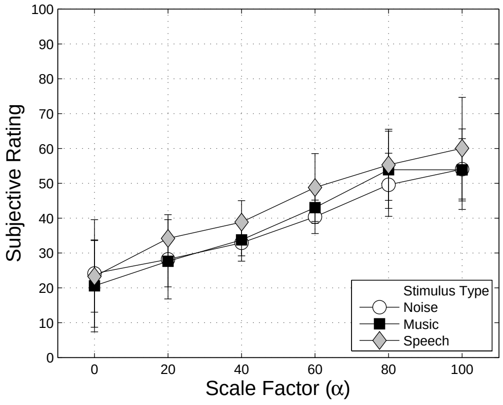


Figure 4: Average externalization ratings pooled over subject and location. Marker color represents base stimulus type. Error bars represent 95% confidence intervals for the pooled data.

no significant main effect for location itself (F(1,1) = 3.943, p = 0.121).

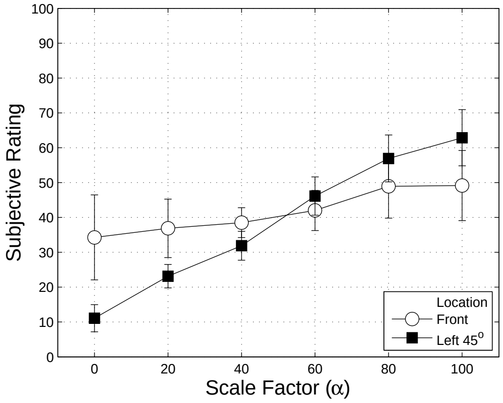


Figure 5: Average externalization ratings pooled over subject and base stimulus type. Marker color represents location. Error bars represent 95% confidence intervals for the pooled data.

## 5. DISCUSSION

The current results agree with the existing literature for the two extreme $\alpha$ conditions. The internalized ratings given to $\alpha = 0$ condition agree with previous studies utilizing only gross binuaral cues, and the full HRTF condition where $\alpha = 100$ showed the expected externalized ratings. More interestingly, the intermediate results suggest that manipulating the strength of narrowband spectral features clearly affects the perceived externalization in a systematic fashion. This implies that the parameterization technique introduced here might be adequate for an externalization or distance based display technology.

Somewhat surprisingly, no effect was found for the different types of base stimuli, despite the fact that they differed significantly in terms of their long-term average spectral profile. Based on those differences the results of Little *et al.* [9] would suggest that the high-frequency roll off seen for the speech and to a lesser extent music would result in higher externalization ratings compared to the flat spectrum noise. This discrepancy could be explained by the blocked nature the experiment since different base stimuli were never compared directly. It is also likely that the high frequency roll off cue is only used when the different stimuli are assumed to be from the same original sound source, a situation clearly not applicable across base stimuli.

The interaction between the $\alpha$ parameter and the stimulus location may be additional evidence to suggest a relationship between externalization and distance perception. By examining the spectral profiles illustrated in Fig. 1, we can clearly see known low-frequency ILD distance cues present for the lateral location that are not available for the front location. The left $45^o$ profiles clearly show an increase in low frequency ILD cues as $\alpha$ is decreased similar to the near-field HRTF cues observed by Brungart *et al.* [14], and an increase in spectral roll off as $\alpha$ is increased similar to the propagation-related distance cue described by Little [9]. The front location only contains the roll off cue due to it's lack of ILD.

In addition to the current positve results, to be valuable as a display technology the parameterization should preserve both the perceived sound quality and localization accuracy. Further research will have to be completed in order to investigate these factors.

## 6. SUMMARY

The current work describes a simple parameterized method for controlling the perceived externalization of a sound source based on flattening of the monaural HRTF spectra. Examinations show that this method can be related to previously observed cues used for auditory distance perception, and a subjective evaluation demonstrates the technique is capable of producing the desired perceptual results.

## 7. REFERENCES

[1] S. Mehrgardt and V. Mellert, “Transformation of the external human ear,” *J. Acoust. Soc. Am.*, vol. 61, pp. 1567–1576, 1977.

[2] A. W. Bronkhorst, “Localization of real and virtual sound sources,” *J. Acoust. Soc. Am.*, vol. 98, pp. 2542–2553, 1995.

[3] D. S. Brungart, G. D. Romigh, and B. D. Simpson, “Rapid collection of HRTFs and comparison to free-field listening,” in *International Workshop on the Principles and Applications of Spatial Hearing*, 2009.

[4] R. Martin, K. McAnally, and M. Senova, “Free-field equivalent localization of virtual audio,” *J. Acoust. Soc. Am.*, vol. 49, pp. 14–22, 2001.

[5] J. Blauert, *Spatial Hearing*. The MIT Press, 1997.

[6] W. M. Hartman and A. Wittenberg, “On the externalization of sound images,” *J. Acoust. Soc. Am.*, vol. 99, pp. 3678–3688, 1996.

[7] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, “The contribution of head movement to the externalization and internalization of sounds,” *PLoS*, vol. 8, p. 1, 2013.

[8] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica*, vol. 91, pp. 409–420, 2005.

[9] A. D. Little, D. H. Mershon, and P. H. Cox, "Spectral content as a cue to percieved auditory distance," *Perception*, vol. 21, pp. 405–416, 1992.

[10] A. W. Mills, "Lateralization of high frequency tones," *J. Acoust. Soc. Am.*, vol. 32, pp. 132–134, 1960.

[11] D. J. Kistler and F. L. . Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, pp. 1637–1647, 1992.

[12] G. A. Soulodre and M. C. Lavoie, "Subjective evaluation of large and small impairments in audio codecs," in *AES International Conference, Florence*, 1999.

[13] *EBU TECH 3253: Sound Quality Assessment Material recordings for subjective tests*.

[14] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources; head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 103, pp. 1465–1479, 1999.