PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Predicting Externalization of Anechoic Sounds

Robert BAUMGARTNER; Piotr MAJDAK

Acoustics Research Institute, Austrian Academy of Sciences, Austria

ABSTRACT

The perceptual association of a sound with a source surrounding the listener is natural but requires complex signal processing to be achieved in virtual or augmented realities presented via headphones. Little is known about the set of spatial features required to elicit an externalized auditory percept, even under anechoic conditions. We investigated the diversity and relevance of different features by conducting a model-based meta-analysis of psychoacoustic studies. As potential features we considered monaural and interaural spectral shapes, spectral and temporal fluctuations of interaural intensity differences, interaural coherence, and broadband inconsistencies (trading) between interaural time and intensity differences in a framework of template-based auditory models. Our model predictions indicate that the monaural spectral shapes and the strength of time-intensity trading present potent cues to explain most previous results.

Keywords: Head-related transfer functions, Virtual auditory reality, Computational modeling

1. INTRODUCTION

Virtual reality systems aim to immerse a listener into a well-externalized three-dimensional auditory space. This requires a perceptually accurate simulation of the listener's natural acoustic exposure. Particularly challenging is to appropriately represent high-frequency spectral cues induced by the pinnae because their morphology is very specific to the listener. To simplify this task, we aimed at developing a phenomenological computational model for sound externalization with a particular focus on spectral cues. The model was designed to predict a listener's degree of externalization based on binaural input signals and the listener's individual head-related transfer functions (HRTFs) under static, anechoic listening conditions.

Previous psychoacoustic experiments assessed sound externalization by subjective distance comparisons (1–7) and/or the discriminability between real and virtual sound sources (5,8). These studies showed that the externalization of anechoic sounds is not affected by broadband approximations of interaural phase differences (5) but may slightly degrade if interaural time delays (ITDs) and interaural level differences (ILDs) are inconsistent (7). The auditory cortex contains an integrated code of sound laterality, but also retains independent information about ITD and ILD cues. This cue-related information might be used to assess how consistent the cues are, and thus, how likely they would have arisen from the same source (9).

While reverberation is essential to accurately estimate the distance of a sound source(10), reverberation alone is not sufficient to externalize a sound (2,6). In order to be externalized, sounds need to provide spectral cues in the direct path (1,6). Spectral cues are most effective in the high frequency range of pinna cues (2) but saliency of spectral cues is also important if sounds do not contain energy at those high frequencies (2,6).

Past studies on spectral cues of spatial hearing mainly focused on sound localization in sagittal planes. Several localization studies suggest that the auditory system processes spectral cues within monaural pathways (reviewed in 11). Animal studies (12) and model simulations (11) provide further evidence that this process might focus on analyzing positive gradients in the stimulus spectrum, but clear human psychoacoustic evidence is still missing and the transferability of findings from directional localization to sound externalization needs to be clarified.

A recent attempt to model the effect of spectral cue saliency on sound externalization compared interaural spectral level differences between the stimulus and internal reference templates (6). However, interaural spectral comparisons are in contradiction to the current understanding of the monaural processing of spectral localization cues and to previous experimental results showing that spectral manipulations degrade externalization also if interaural differences are preserved (5).

By conducting a model-based meta-analysis, this study aims to clarify the auditory mechanisms of spectral cue processing for auditory externalization and, in particular, the perceptual weighting of spectral cues in comparison to other spatial auditory cues.

2. METHODS

2.1 Model

The model architecture generally follows a template-matching procedure applied to a set of different cues (Figure 1). The model evaluates positive gradients in the spectral shape of magnitude profiles (MSG; c.f., 13), the spectral shape of ILDs (ISS; c.f., 6), the spectral standard deviation of ILDs (ISSD; c.f., 14), the inconsistency between ITD and ILD (ITIT), and the interaural coherence (IC) as summarized in Table 1. As a control cue, we also evaluated the difference in overall sound pressure level between target and template stimuli considering that loudness differences between the consecutively presented stimuli may have affected the perceptual ratings.

Table 1 – Externalization cues evaluated within the model

Cue	Description
MSG	Monaural spectral gradients (c.f., 11,13)
ISS	Interaural spectral shape (c.f., 6)
ISSD	Interaural spectral standard deviation (c.f., 14)
IC	Interaural coherence (c.f., 15)
ITIT	Interaural time-intensity trading (ITD vs. ILD)
SPL	Overall level difference between target and reference stimulus

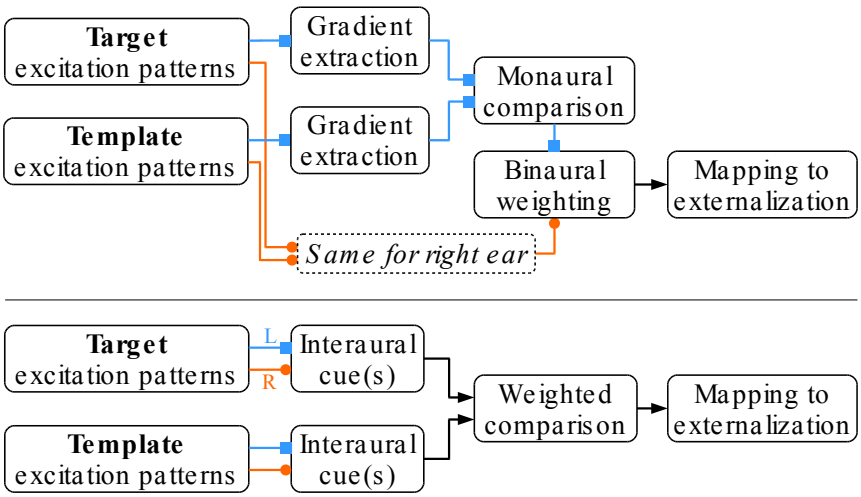

Figure 1 – Investigated structures of externalization models based on monaural (top) and/or interaural (bottom) cue template comparisons

ITDs were derived from binaural signals low-pass filtered at 3 kHz by extracting the time lag that yields maximum IC of the temporal energy envelope ("MaxIACCe lp" in 16). Overall level differences denoted the difference in dB RMS levels between the target and reference stimuli, thresholded by a just-noticeable difference (JND) of 1 dB and averaged across ears.

All other cues were evaluated after filtering the target and template signals through a bank of fourth-order Gammatone filters with a regular spacing of one equivalent rectangular bandwidth. We computed the logarithm of the root-mean-square (RMS) energy within every frequency band to obtain spectral excitation profiles (c.f., 13). Audibility thresholds for band-limited signals were approximated by generally considering stimulus sound pressure levels of 70 dB and a within-band threshold of 20 dB. To allow the evaluation of temporal fluctuations, the profiles were evaluated in non-overlapping blocks of 5 ms, which has been considered as the integration time for spectral cues (17,18). Assuming stationary input signals, the spectral profiles were averaged over time.

The IC metric was denoted as the IC difference between target and reference:

$$d_{IC} = IC_T - IC_R \text{ with } IC = \lim\inf_{\tau} \frac{\int x_L(t-\tau)x_R(t)dt}{\sqrt{\int x_L(t)^2 x_R(t)^2 dt}} \text{ and } \tau \in [-1,1] \text{ ms.} \quad (1)$$

For the ISSD metric, the model evaluated standard deviations of ILDs across frequency bands, computed the absolute difference of these deviations between the target and template, relative to the template deviation:

$$d_{ISSD} = 1 - \frac{SD_f(ILD_T(f))}{SD_f(ILD_R(f))}. \quad (2)$$

For the ISS metric, absolute differences between the target and template ILDs were evaluated, then differences smaller than 1 dB, considering JNDs of ILDs to exceed this range (Mills, 1960), were set to zero and remaining differences were normalized by the template ILDs and finally averaged across frequency bands:

$$d_{ISS} = \frac{1}{N_f}\sum_f \left|\frac{ILD_T(f) - ILD_R(f)}{ILD_R(f)}\right|. \quad (3)$$

The ITIT was denoted as the difference magnitude between target-to-template ratios of ITD and ILD (averaged across frequency bands):

$$d_{ITIT} = \left|\frac{\Delta ITD_T}{ITD_R} - \frac{\Delta ILD_T}{ILD_R}\right|, \quad (4)$$

with $\Delta ITD_T = ITD_T - ITD_R$ of 0 if smaller than 20 µs, $\Delta ILD_T = ILD_T - ILD_R$ of 0 if smaller than 1 dB.

For the MSG, positive spectral gradient profiles were obtained by differentiating the excitation profiles ($p(f) \rightarrow p'(f)$) and softly restricting the value range by an elevated arctangent as proposed in (13):

$$MSG = arctan\left(p'(f) - \frac{\pi}{2}\right) + \frac{\pi}{2}. \quad (5)$$

These gradient profiles were then compared between the target and template separately for each ear by applying the same procedure as for the ISS metric, that is, calculating absolute target-to-template differences, normalizing differences larger than 1 dB by the template gradients, and finally averaging those differences across frequencies. The MSG distance metrics for the two ears were then combined according to the binaural weighting function defined in (11), effectively increasing the perceptual weight of the ipsilateral ear with increasing lateral eccentricity:

$$d_{MSG} = \frac{d_{MSG,l} - d_{MSG,r}}{1 + e^{-\phi/\Phi}} + 1, \quad (6)$$

with $\phi \in [-90°, 90°]$ denoting the lateral angle (left is positive) and $\Phi = 13°$.

Similarly for all metrics, an exponentially decaying mapping function with the exponent reciprocally scaled by the sensitivity parameter S_{cue} was used to map cue-specific deviations d_{cue} from the template to externalization scores E_{cue}.

2.2 Assessment

Prediction errors are defined as the RMS of differences between actual and predicted externalization score ratings normalized by the range of the rating scale used in the individual study. Optimal cue-specific sensitivity parameters are obtained by minimizing the prediction error individually for each study. The model's performance is assessed for various conditions collected from a representative set of experiments (1,2,5,6).

3. Results

3.1 Effects of low-frequency alterations

Hartmann and Wittenberg (5) synthesized the vowel */a/* with a tone complex consisting of 38 harmonics of the fundamental frequency of 125 Hz, yielding a sound bandlimited up to 4750 Hz. This sound was presented via headphones and filtered with individualized HRTFs. In one experiment, the magnitudes of all harmonics up to a certain harmonic *n'* were set to the interaural

average, effectively removing ILDs up to that harmonic's frequency. In the other experiment, the ipsilateral magnitude spectrum was flattened up to *n'* while the contralateral magnitudes were shifted in parallel, effectively maintaining the original ILDs but changing the monaural spectral profiles. In both experiments, the listeners were asked to rate the degree of auditory externalization on a continuous metric scale with minimum values referring to inside-the-head localization and maximum values referring to localization at the actual loudspeaker position.

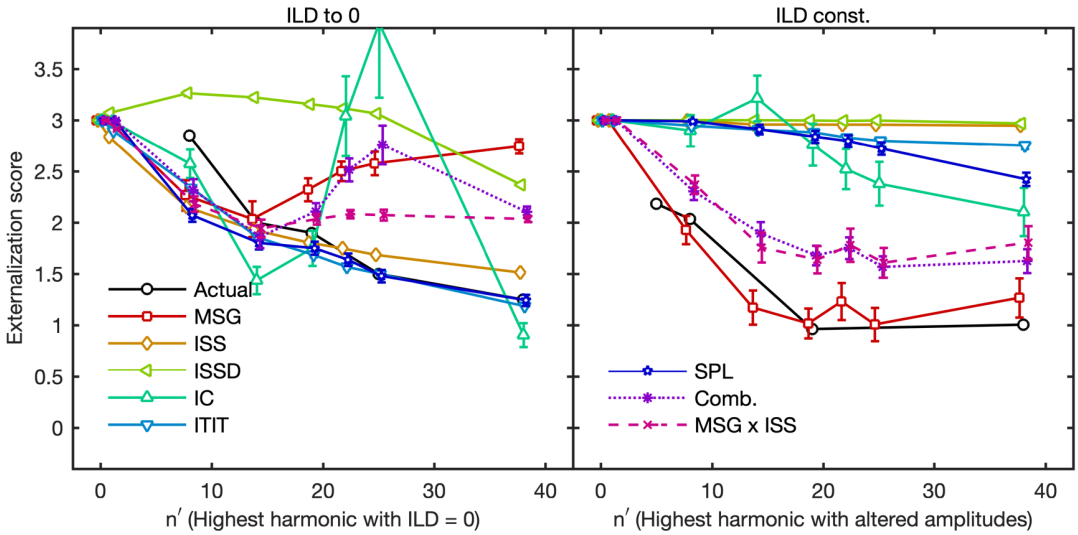

Figure 2 – Model simulations to predict the effect of low-frequency alterations up to a certain harmonic of a complex tone with a fundamental frequency of 125 Hz. Actual data from Figures 7 and 8 from (5)

The results in Figure 2 (left) show that externalization degradation induced by ILDs successively set to zero were well predicted by the ISS, ITIT, and SPL cues. The more harmonics at which ILDs have been removed, the more the target ISS is flattened and deviates from the template ISS, the smaller is the average ILD in comparison to the constant ITD, and the higher is the SPL. The ISSD cue leads to non-monotonic predictions that strongly depend on the center frequencies of peaks in the spectral shape of the template ILDs. IC differences are very small and thus noisy. The MSG cue is rather insensitive to the changes in ILDs because at low frequencies the complex acoustic filtering of the pinnae is negligible and thus monaural spectral shapes are quite similar at both ears and only marginally affected by interaural averaging.

The degradation induced by flattening the ipsilateral spectrum while maintaining the original ILDs (Figure 2, right) was well predicted by the MSG cue only. Predictions based on interaural cues perform poorly because the experimental condition is designed to not affect these cues.

Overall, the MSG cue yields the smallest prediction errors across both experiments, whereby there is a marked difference in predictive power between the two different experimental manipulations. The interaction term comprising MSG and ITIT performs best overall.

3.2 Effects of spectral smoothing

Hassager et al. (6) presented Gaussian white noise bandlimited from 50 to 6000 Hz. These sounds were filtered with individualized BRIRs (RT60 between 300 and 600 ms) in order to simulate sound sources positioned at azimuths of 0° and 50°. As independent experimental variable, Gammatone filters with various equivalent rectangular bandwidths (ERBs) were used to spectrally smooth the direct path portion (until 3.8 ms) of the BRIRs. Filters with larger ERBs more strongly smoothed the shape of the magnitude spectrum. Similar to the previous study (5), listeners rated auditory externalization on a continuous scale.

Model simulations were based on (anechoic) HRTFs because the original BRIRs were not accessible. This is not critical assuming that only the direct path is relevant if also only the direct path has been modified during the experiment. Despite the non-monotonicities of the ISSD cue and the minor changes in IC and SPL, most predictions follow the systematic trend of externalization degradation with increasing bandwidth factor (see Figure 3). Moreover, all these cues are consistent with the actual results in that they are insensitive to spectral smoothing below one ERB. Overall, the actual results are best predicted on the basis of the MSG cue.

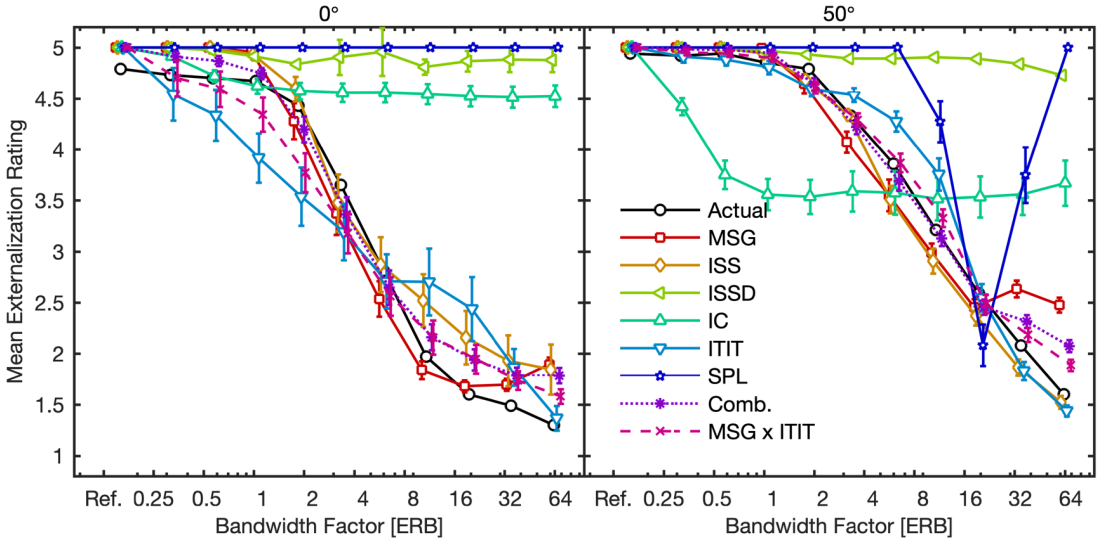

Figure 3 – Model simulations to predict the effect of spectral smoothing on the externalization of sounds from various azimuths (0° and 50°). Actual data represent direct-sound condition from (6)

Baumgartner et al. (1) also tested 15 normal-hearing listeners on the effect of spectral smoothing but focusing on the high-frequency range between 1 to 16 kHz where the pinna induces the most significant directional spectral variations. In contrast to the other studies, listeners judged auditory externalization not absolutely but relatively within paired comparisons. Absolute externalization scores were then estimated from the paired judgements via probabilistic model fitting (19). Only scale estimates yielding fitting statistics within a heuristic range of $0.001 < \chi^2 < 10$were considered trustworthy and used for evaluation. Scale estimates for six out of twelve listeners fulfilled this requirement.

Model predictions were based on listener-specific HRTFs and also assessed against listener-specific results. As a consequence of large inter-individual differences the errors are comparably high, while the average predictions shown in Figure 4 are still similar to the actual results. The IC and SPL cues yield particularly poor results, whereas all other cues perform quite similarly across the limited set of experimental conditions.

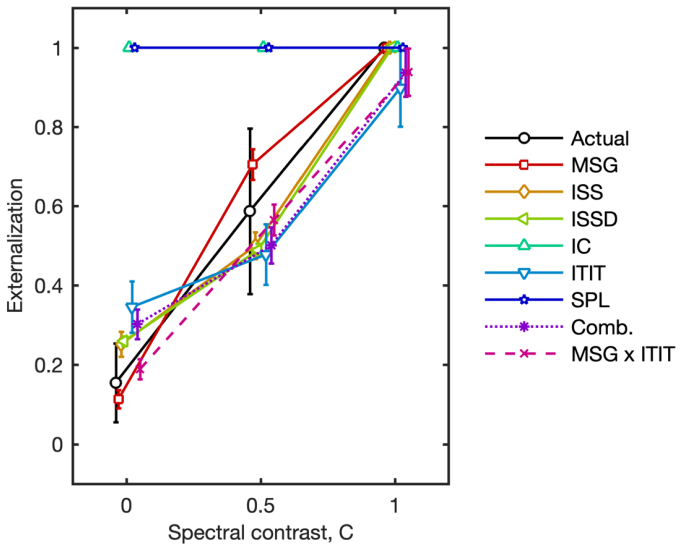

Figure 4 – Model simulations to predict the effect of spectral smoothing on the externalization of sounds from various azimuths (±90° or 0°). Actual externalization scores estimated from paired-comparison data from (1)

3.3 Effects of binaural microphone casing and stimulus bandwidth

Boyd et al. (2) presented broadband speech samples and used individualized BRIRs to simulate a talker positioned at 30° azimuth. They compared externalization ratings for in-the-ear (ITE) and behind-the-ear (BTE) microphone casings as well as broadband (BB) and 6.5-kHz-low-pass (LP) filtered stimuli at various mixing ratios with stereophonic recordings providing only an ITD. The amount of reverberation remained constant across mixing ratios.

For model simulations, original BRIRs were only available for 3 out of 7 (normal-hearing) listeners. As suggested by earlier modeling work (6), spectral-shape cues (MSG and ISS) were only evaluated for the direct path component of the BRIRs. Hence, for MSG and ISS, only the first 5 ms of the BRIRs were used, faded-out by a cosine taper of 1 ms ending at the 5 ms. The simulation results in Figure 5 show that all but the SPL and ITIT cues performed quite well.

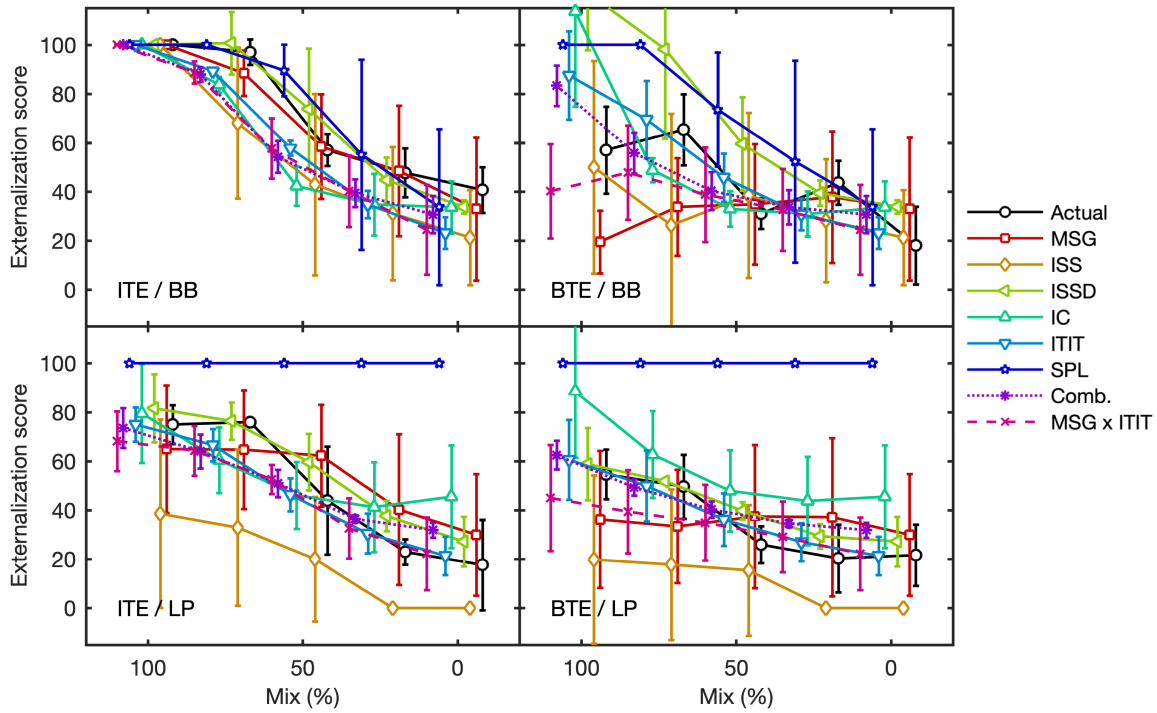

Figure 5 – Model simulations to predict the effects of microphone casing and stimulus bandwidth for various mixes between simulations based on individualized acoustics (100%) and ITDs only (0%). Actual data from (2) represents normal-hearing condition with one talker at 30° azimuth and reverberation. ITE: in-the-ear casing; BTE: behind-the-ear casing; BB: broadband stimulus; LP: low-pass filtered at 6.5 kHz

4. DISCUSSION

Model simulations were conducted for three studies using (anechoic) HRTFs and one study using (reverberant) BRIRs. In all anechoic conditions, predictions based on either MSGs or inconsistencies between ITD and ILD (ITITs) performed very well. With results for high-pass filtered stimuli excluded, the overall picture for anechoic conditions seems to suggest that listeners base their externalization ratings on the one of the two cues that indicates largest deviations from the reference sound. Under reverberant conditions, all interaural cues (ISS, ISSD, IC, ITIT) yielded similarly small prediction errors, slightly favoring also the ITIT cue.

The sensitivity parameters used to scale the mapping function from cues to externalization scores were optimized separately for every cue and study. Separate optimization is reasoned by the limited quantitative comparability of subjective externalization ratings due to differently trained subjects, different contexts, different experimental procedures especially with respect to presenting reference stimuli, and many other methodological differences. Nevertheless, the optimization procedure yielded similar sensitivity parameters for the favorite cues (MSG and ITIT) across studies, which underlines the generalizability of our results.

In all simulations template comparisons were only performed for the target direction, ignoring that there could be a strong match to a template from another direction yielding strong externalization. We did that because of two reasons. First, except for one study (1), listeners in all the other experimental paradigms were asked to rate externalization against a fixed reference stimulus. Hence, we assumed these ratings to be biased by the fixed reference location even though it is hard to estimate how much these ratings were affected by directional localization changes, for instance, within sagittal planes because of spectral cue changes. Second, listener-specific BRIRs for source directions all around the listeners were not available.

5. CONCLUSIONS

The present investigations suggest that monaural spectral-shape cues are important for sound externalization especially under anechoic listening conditions. However, the actual nature and processing of these cues is still poorly understood. The MSG cue implemented here has only been motivated by physiological findings in cats and psychoacoustic model simulations in the context of sagittal-plane sound localization. Future experiments could be targeted to clearly dissociated the relevance of positive vs. negative spectral gradients. Moreover, investigating this dissociation in combination with a systematic variation of the amount of reverberation would allow to evaluate the transition in perceptual relevance from MSG and ITIT under anechoic conditions to a potentially broader set of cues under reverberant conditions.

ACKNOWLEDGMENTS

We thank Bill Withmer for kindly providing the original data from Boyd et al. (2). This work was supported by the Austrian Science Fund (FWF J 3803-N30) and Facebook Reality Labs.

REFERENCES

1. Baumgartner R, Reed DK, Tóth B, Best V, Majdak P, Colburn HS, et al. Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias. Proc Natl Acad Sci. 2017 Sep 5;114(36):9743–8.
2. Boyd AW, Whitmer WM, Soraghan JJ, Akeroyd MA. Auditory externalization in hearing-impaired listeners: The effect of pinna cues and number of talkers. J Acoust Soc Am. 2012 Mar;131(3):EL268–74.
3. Catic J, Santurette S, Buchholz JM, Gran F, Dau T. The effect of interaural-level-difference fluctuations on the externalization of sound. J Acoust Soc Am. 2013 Aug 1;134(2):1232–41.
4. Catic J, Santurette S, Dau T. The role of reverberation-related binaural cues in the externalization of speech. J Acoust Soc Am. 2015 Aug 1;138(2):1154–67.
5. Hartmann WM, Wittenberg A. On the externalization of sound images. J Acoust Soc Am. 1996 Jun;99(6):3678–88.
6. Hassager HG, Gran F, Dau T. The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment. J Acoust Soc Am. 2016 May 1;139(5):2992–3000.
7. Zhang PX, Hartmann WM. On the ability of human listeners to distinguish between front and back. Hear Res. 2010 Feb;260(1–2):30–46.
8. Brimijoin WO, Boyd AW, Akeroyd MA. The Contribution of Head Movement to the Externalization and Internalization of Sounds. PLoS ONE. 2013 Dec 2;8(12):e83068.
9. Edmonds BA, Krumbholz K. Are Interaural Time and Level Differences Represented by Independent or Integrated Codes in the Human Auditory Cortex? J Assoc Res Otolaryngol. 2014 Feb 1;15(1):103–14.
10. Kolarik AJ, Moore BCJ, Zahorik P, Cirstea S, Pardhan S. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. Atten Percept Psychophys. 2016 Feb 1;78(2):373–95.
11. Baumgartner R, Majdak P, Laback B. Modeling sound-source localization in sagittal planes for human listeners. J Acoust Soc Am. 2014 Aug;136(2):791–802.
12. Reiss LAJ, Young ED. Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus. J Neurosci. 2005 Feb;25(14):3680–91.
13. Baumgartner R, Majdak P, Laback B. Modeling the Effects of Sensorineural Hearing Loss on Sound Localization in the Median Plane. Trends Hear. 2016;20:1–11.
14. Georganti E, May T, Par S van de, Mourjopoulos J. Sound Source Distance Estimation in Rooms based on Statistical Properties of Binaural Signals. IEEE Trans Audio Speech Lang Process. 2013 Aug;21(8):1727–41.
15. Hassager HG, Wiinberg A, Dau T. Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment. J Acoust Soc Am. 2017 Apr 1;141(4):2556–68.
16. Katz BFG, Noisternig M. A comparative study of interaural time delay estimation methods. J Acoust

Soc Am. 2014 Jun 1;135(6):3530–40.
17. Hartmann WM, Best V, Leung J, Carlile S. Phase effects on the perceived elevation of complex tones. J Acoust Soc Am. 2010;127 (5):3060–72.
18. Hofman PM, Opstal AJV. Spectro-temporal factors in two-dimensional human sound localization. J Acoust Soc Am. 1998 May;103(5):2634–48.
19. Wickelmaier F, Schmid C. A Matlab function to estimate choice model parameters from paired-comparison data. Behav Res Methods Instrum Comput. 2004;36(1):29–40.