Audio Engineering Society
# Conference Paper
Presented at the Conference on
Audio for Virtual and Augmented Reality
2016 Sept 30 – Oct 1, Los Angeles, CA, USA



## Perceptual weighting of binaural information: toward an auditory perceptual "spatial codec" for auditory augmented reality

G. Christopher Stecker[1] and Anna C. Diedesch[1,2,3]

[1]*Department of Hearing and Speech Sciences, Vanderbilt University School of Medicine*
[2]*Oregon Hearing Research Center, Oregon Health & Science University*
[3]*National Center for Rehabilitative Auditory Research, Portland VA Medical Center*

Correspondence should be addressed to Chris Stecker (cstecker@spatialhearing.org)

### ABSTRACT

Auditory augmented reality (AR) requires accurate estimation of spatial information conveyed in the natural scene, coupled with accurate spatial synthesis of virtual sounds to be integrated within it. Solutions to both problems should consider the capabilities and limitations of the human binaural system, in order to maximize relevant over distracting acoustic information and enhance perceptual integration across AR layers. Recent studies have measured how human listeners integrate spatial information across multiple conflicting cues, revealing patterns of "perceptual weighting" that sample the auditory scene in a robust but spectrotemporally sparse manner. Such patterns can be exploited for binaural analysis and synthesis, much as time-frequency masking patterns are exploited by perceptual audio codecs, to improve efficiency and enhance perceptual integration.

## 1 Introduction

Perceptual audio coding has been tremendously successful in a wide range of audio applications requiring significant data compression with minimal perceptual impact. Psychoacoustic codecs (e.g., AAC, AC-3, and MP3) provide not only smaller file sizes and lower bit rates for digital audio, but also support higher channel counts and more detailed scene-based audio representations. The technological development of these algorithms dates back roughly 30 years, but the psychoacoustical foundation of that work dates back much further. Beginning in the 1930s, Harvey Fletcher and colleagues at Bell Labs worked to quantify spectral and temporal patterns of auditory excitation and masking in the ear [7]. Excitation patterns, or masking patterns, describe how sounds are represented in cochlear activity and how these representations interact when multiple sounds are presented simultaneously. Thus, the patterns can be used to identify the components of a signal that are most and least important to auditory perception due to such interactions. A major insight is that unmasked, audible, components are sparsely distributed in time

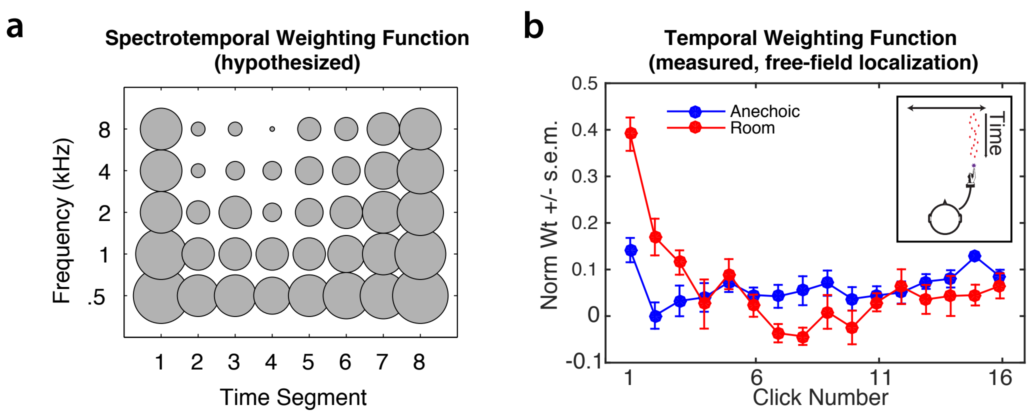


**Fig. 1:** Illustration of spatial cue weighting functions. In (a), a brief stimulus spans a range of frequency components (0.5–8 kHz, vertical axis) and a temporal duration divided into eight segments for plotting (horizontal axis). The relative impact of spatial cues in each spectrotemporal bin is plotted by the size of the circles and comprises the "spectrotemporal weighting function," or STWF. The illustration combines three features established by prior research: (1) dominance of the sound onset [1, , see (b)], (2) upweighting of ILD cues near sound offset [2, 3], and (3) dominance of ITD cues in low-frequency components [4, 5]. In (b), temporal weighting functions (TWF) are plotted for localization of narrowband click trains (4 kHz center frequency, repeating at 5-ms ICI) in the free-field. Weights were normalized to sum to 1.0 across all 16 clicks and plotted against click number. Error bars indicate standard error of the mean across 5 listeners. Onset dominance and upweighting are both clearly present in anechoic TWFs (blue lines). Red lines indicate stimuli presented in a 10 m by 10 m room, simulated using the image method [6] with four side walls ($\alpha = 0.5$) and reflections up to 13th order. The room condition dramatically enhanced the weight applied to onset cues, consistent with the importance of onsets in reverberant listening.

and frequency; much of the remaining sound can be discarded without major perceptual effects. Algorithms that exploit that knowledge can achieve dramatic compression with minimal perceptual effects.

This paper explores an analogous phenomenon in the spatial domain, specifically the perceptual weighting of auditory spatial cues by human listeners. Like masking patterns, the patterns of spatial cue weighting can be used to identify the most and least important spatial features of sounds. Recent evidence suggests that much as in masking patterns, the most relevant spatial features of sounds are distributed rather sparsely in both time and frequency [8]. Exploiting this knowledge could lead to algorithms for data compression in spatial audio synthesis and to improvements in spatial audio analysis, two areas of particular relevance to audio applications in virtual reality (VR) and augmented reality (AR).

Spatial psychoacoustics provides a clear picture of the acoustical cues that drive listeners' perception of auditory space. These include interaural differences of time (ITD) and level (ILD) as well as monaural spectral features. All of these are captured by the linear time-invariant features of the (anechoic) HRTF. In fact, the HRTF provides a reliable-seeming "lookup table" for these cues: at any given frequency, each possible sound-source location gives rise to a particular combination of ITD, ILD, and intensity values. One might therefore consider the various cues to be *in agreement* with one another and with the sound-source location. This situation holds, however, only for isolated sources presented in anechoic conditions. Competing sounds, echoes, and reverberation profoundly alter these cues, and they do so in a cue-, frequency-, and time-dependent manner. Thus, in real-world listening cues often disagree, and there is no simple correspondence between cue values and source locations. Perceptual weighting studies reveal that the human auditory system encodes spatial information in a way that is robust to these types of effects.

In a typical reverberant space, direct sound is followed by a series of early reflections and finally a diffuse reverberant sound field. Thus, the very earliest arriving sound (the "first wavefront" [9]) carries mainly direct (source) cues, whereas the ITD, ILD, and spectral cues of later-arriving sound are distorted by room effects. Specifically, ITD cues fluctuate erratically over time, and ILD cues diminish toward zero, in reverberation [10]. Depending on the perceptual correlates of these changes, listeners' perception in reverberant space ought to evolve over time, and sound sources should appear to move around. This is not the case for normal perception in reverberant environments, however; rather, listeners perceive a stable auditory scene with sources that remain fixed unless in actually motion. How is this accomplished?

One answer is that the brain emphasizes the most reliable spatial cues: ITD and ILD at onset, ILD at offset, and so forth. The brain discounts cues that are likely to be impacted negatively by echoes and reverberation. This differential weighting of auditory spatial cues provides a form of perceptual constancy for auditory space. For weighting of cues over time, the process can be characterized by "temporal weighting functions" (TWF) [11, 12, 13]. Similarly, binaural cues also differ in reliability across frequency, so there is an equivalent set of spectral weighting functions. We can put the two together in the form of spectro-temporal weighting functions for binaural-cue sensitivity. An illustration appears in Fig. 1a.

Audio rendering for VR has typically attempted to accurately reproduce the *physical* features of spatial audio, for example by capturing HRTFs and binaural room impulse responses (BRIR). The goal is to present the "correct" acoustical cues—with equal weight—at the listener's ear, under the assumption that correct cues will result in realistic perception of the virtual auditory scene. Our current understanding of spatial cue weighting, however, suggests an alternate approach that maximizes the fidelity of highly weighted spatial cues at the expense of weakly weighted ones.

An additional constraint in AR but not VR applications is the need to seamlessly integrate spatial information across multiple layers of synthetic and natural sound. Satisfying that constraint will require systems that can track and predict how listeners perceive a given scene in order to synthesize new layers in a perceptually consistent manner. Psychoacoustic models of spatial cue weighting will significantly benefit this process by emphasizing the most robust and critical spatial features of sound.

Making use of perceptual weighting for applications in audio for VR and AR requires two steps: (1) the estimation of spatial cue weighting functions and (2) the application of those functions to spatial sound synthesis. In the next sections, we describe examples of each.

## 2 Estimation of spatial cue weighting functions

Work in our laboratory has focused on quantifying cue-weighting across spatial cue type [e.g. ITD versus ILD 14], across frequency, and particularly over time. Here, we will focus on the latter as an example of both the estimation of spatial cue weighting and its use to guide spatial audio synthesis.

Numerous studies have used statistical regression methods to recover TWFs for auditory spatial cues (e.g., [11, 12, 13]). In this approach, listeners make spatial judgments of sounds whose spatial cues vary randomly and by a small amount over time. An example is illustrated in the inset of Fig. 1b. The vertical dimension represents time and the horizontal dimension represents stimulus location in azimuth. The red dots indicate brief segments of a sound, each of which is presented from a slightly different location. The drawing illustrates a single trial; other trials would present a different set of random spatial variations, and a different overall azimuth.

For such sounds containing multiple temporal segments $i$, TWFs can be estimated by regression of listeners' localization responses $\theta_R$ onto the independent segment locations $\theta_i$:

$$\hat{\theta}_R = \sum_{i=1}^{16} \beta_i \theta_i + k, \tag{1}$$

Relative weights $w_i$, which make up the TWF, are typically estimated after normalizing the regression coefficients $\beta_i$ over the stimulus duration (the bias term $k$ is typically not included in the TWF):

$$w_i = \frac{\beta_i}{\sum_{j=1}^{16} \beta_j}. \tag{2}$$

The weights $w_i$ thus quantify the relative influence of binaural cues in each temporal segment on the listeners' spatial judgments.

Many past studies have estimated TWFs using trains of rapidly presented clicks. Such stimuli are simple to temporally segment (i.e. each click is a segment), and gaps between clicks help to minimize acoustical interactions between successive segments. Fig. 1b plots TWFs ($w_i$ versus time) for brief trains of filtered clicks presented at 5-ms interclick interval (ICI) from a 360° array of 64 ear-height loudspeakers in the Vanderbilt Bill Wilkerson Anechoic Chamber Laboratory. Stimuli varied in azimuth across a range of ±56.25°, and individual clicks were perturbed up to ±11.25° by selecting a loudspeaker for each click.

When the experiment was conducted in simple anechoic space, TWFs (blue line in Fig. 1b) revealed clear onset dominance and "upweighting" of later clicks [2], two features that have been repeatedly demonstrated in localization tasks of this type [12, 3, 13]. The experiment was repeated in a virtual room condition. The image method [6] was used to simulate a 10 m by 10 m room with four lateral walls (front, back, and two sides, all $\alpha = 0.5$). Each simulated reflection—up to 13th order—was presented at the correct azimuth (rounded to the nearest loudspeaker), intensity, and delay for the calculated path. Reverberation time T60 was 0.3 s. TWFs measured in the simulated room (red lines) revealed significantly stronger onset dominance, consistent with the expected immunity of onset cues to distortion by echoes.

The TWFs plotted in Fig. 1b and in past studies illustrate that listeners' spatial impression is dominated by a subset of the potentially available spatial cues, namely the cues available at sound onset and in some cases offset [13]. In particular, the middle portion of a brief sound appears to have nearly zero influence on listeners' spatial impression. This suggests that for simple sounds like these, manipulation of the onset alone (or onset+offset) should be sufficient to control spatial perception. But what about more complex sounds such as noises or sounds with modulated envelopes? A separate experiment measured TWFs for modulated noises (Fig. 2).

## 3 TWF for amplitude-modulated sounds

As illustrated in Fig. 2a, stimuli were trains of 1-ms white-noise bursts, repeating at a rate of 500 Hz (i.e., 2-ms ICI). Stimuli were delivered over headphones, with ITD and ILD manipulated in a correlated fashion across a range of ±600 $\mu s$ and ±6 dB. Thus, values of $\theta_i$ in Eq. 1 were ITD/ILD combinations rather than loudspeaker azimuths. On each trial, listeners indicated the perceived lateral position ($\theta_R$) on a touchscreen monitor. In one condition, stimuli were presented with a flat amplitude envelope (i.e., all noisebursts were presented at equal intensity). In other conditions, sinusoidal amplitude modulation (AM) was applied at three rates resulting in one, two, or four AM cycles over the stimulus duration. TWFs obtained in all four conditions are plotted in Fig. 2b. These reveal clear onset dominance in the flat condition. Note that the greater onset dominance compared to Fig. 1b is consistent with previous studies of TWF at 2-ms ICI [12]. For AM conditions, the largest weights occurred during the earliest rising part of each modulation period. Note that for all three AM conditions, "click 1" was actually silent; the largest weights were always obtained on the very first non-silent click (click 2) despite its low amplitude. In contrast, the most intense clicks received very low weights, as did clicks aligned with the falling phase of the envelope. These results are consistent with the dominance of the overall onset in sounds with flat envelopes, and with the importance of rising envelopes for ITD processing at high [16] and low [17] frequencies.

The results of TWF measurements as illustrated in Figs. 1 and 2 reveal consistent patterns in listeners' weighting of binaural and spatial information over the durations of brief sounds. Such patterns could be exploited in "binaural listening" algorithms that attempt to estimate listeners' spatial perception of real-world sound, for example to align AR content to that perception. The patterns could also be used to guide spatial audio synthesis, a problem that is taken up in the following section.

## 4 Weighting-guided synthesis of spatial audio

Spatial cue weighting patterns (e.g., TWFs) may be exploited similarly to masking patterns that reflect cochlear excitation [7]. In the same way that masking patterns suggest which components of a sound will have the most impact (and which the least) on audibility and sound quality, the TWFs suggest which temporal segments of a sound will have the most impact (and which the least) on spatial impression. Specifically, TWF measurements suggest that sound localization is dominated by spatial cues present during rising envelope slopes. A rather straightforward test of that idea is to resynthesize sounds with different spatial information in the rising versus falling envelope segments. Fig. 3 illustrates such a test.

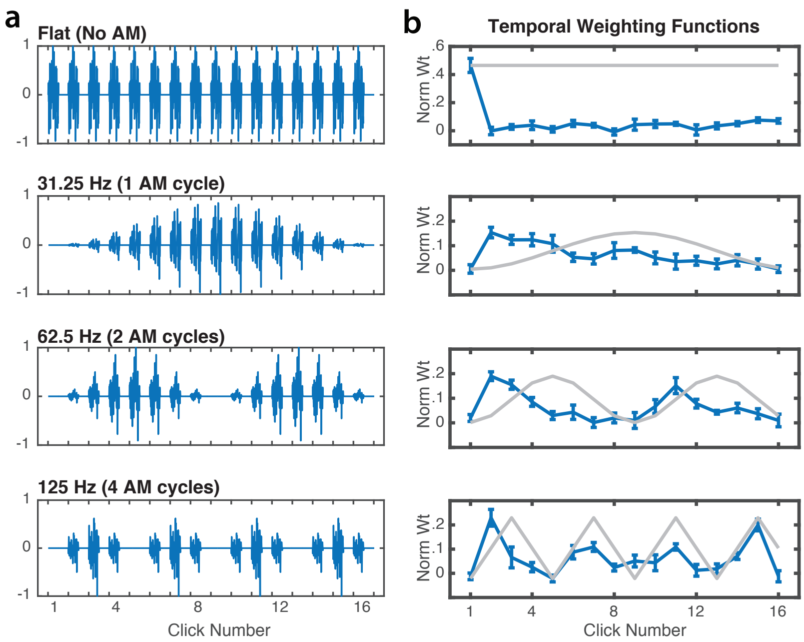


**Fig. 2:** TWFs estimated from amplitude-modulated noises. Stimuli (a) were trains of 1-ms noise bursts presented at 2-ms ICI. In three conditions, sinusoidal amplitude-modulation (SAM) envelopes were applied with a frequency of 31.25, 62.5, or 125 Hz. Each of these conditions, along with an unmodulated control condition, is illustrated in a separate row of panels. Panels in (b) plot TWFs for obtained from regression of lateralization judgments on ITD/ILD cue values (mean of 8 subjects, ±1 standard error). Gray lines plot the AM envelopes for reference. Onset dominance is clearly evident in the TWF obtained with the unmodulated stimulus. A similar result occurred for AM stimuli: the strongest weights were obtained during the earliest part of the rising SAM envelope (e.g., click 2) in each case. Figure adapted from [15].

Single-syllable words were processed using a 4-channel click-train vocoder (Fig. 3a). The vocoder extracted the speech envelope in each of four frequency bands, centered at 1, 2, 4, and 8 kHz. The four envelopes were then used to modulate the amplitudes click trains (5-ms ICI) that matched the filter in spectral frequency. The envelopes were also used to label clicks according to the positive (rising, blue) or negative (falling, orange) slope of the envelope. Stimuli were presented to listeners in the Anechoic Chamber Laboratory, with rising and falling clicks directed to different loudspeakers separated by 11.25° (Fig. 3b). The overall stimulus location varied from trial to trial over a range of ±45°. Listeners indicated the perceived location of each stimulus using a touch display (Apple iPad).

Judgment data were converted into weights for the rising-envelope and falling-envelope cues using multiple linear regression:

$$\hat{\theta}_R = \beta_{rise}\theta_{rise} + \beta_{fall}\theta_{fall} + k, \quad (3)$$

where $\theta_R$ indicates response azimuth as in Eq. 1, $\theta_{rise}$ indicates the azimuth of the loudspeaker active during the rising envelope, and $\theta_{fall}$ the azimuth during the falling envelope. Rise and fall weights ($\beta_{rise}$ and $\beta_{fall}$) are plotted in Fig. 3c.

Results support the hypothesis that localization of complex sounds is dominated by the spatial cues present during rising envelope slopes. They suggest the sufficiency of spatial audio synthesis that manipulates only the spatial cues carried by rising-envelope clicks. Other clicks can be rendered at 0° (or with uninformative binaural cues) with limited impact on spatial perception: images would appear laterally compressed to roughly 70+% of full-cue lateralization, but at a potentially significant savings in spatial data. Bear in mind that TWFs

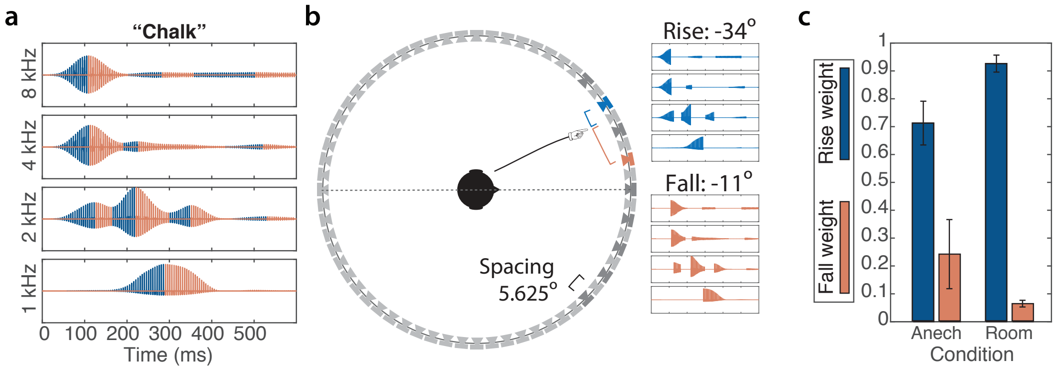


**Fig. 3:** Illustration of TWF-guided spatial synthesis. Single-syllable words (e.g. "chalk") were processed by a 4-channel click-train vocoder (a). Amplitude envelopes in each frequency channel were used to separate each narrowband stimulus into rising-envelope (blue) and falling-envelope (orange) segments. Sounds were delivered via anechoic 64-channel loudspeaker array (b), with rise and fall segments delivered to different loudspeakers. Listeners indicated the perceived location of each presentation. Statistical regression of judgments on loudspeaker locations was used to derive localization weights for rising-envelope (blue) and falling-envelope (orange) segments (c; bars indicate mean $\pm 1$ standard error across 3 listeners). Weighting analysis indicates that spatial perception was strongly dominated ($> 3:1$) by rising-envelope segments. When stimuli were presented in a simulated 10 m by 10 m room (as in Fig. 1b), dominance of the rising-envelope cues was even stronger ($> 10:1$). Figure adapted from [15].

suggest restricting the cues even further; the largest weights appear tightly clustered in the earliest part of each rising envelope rather than distributed throughout it. The current result is only the first of many attempts to specify this more precisely.

When the same experiment was repeated in the simulated room described in the previous section, the rising-envelope cues became even more important (Fig. 3c). That result is consistent with the effects of room simulation on TWFs in Fig. 2, and suggests greater flexibility in future applications to VR and AR when simulated or real reverberation is included.

## 5 Discussion

This paper presents an overview of spatial-cue weighting by human listeners, focusing particularly on temporal weighting as an example. The literature on this topic suggests that strongly-weighted spatial features are distributed rather sparsely within most stimuli. That observation implies a significant potential benefit for spatial analysis and synthesis of audio for VR and AR: namely, that psychoacoustical models can be used to identify and target the most important spatial features of sound. We have argued that such models be used similarly to psychoacoustical models of cochlear excitation and masking: for perceptual coding of compressed spatial data and generation of robust perceptual descriptions of spatial audio elements.

Weighting-guided analysis and synthesis of spatial audio could have numerous applications in entertainment (surround sound for 3D video), VR (data compression for rich spatial scenes), and AR (analysis and matching of synthetic to natural source material). Particularly when applied to AR, these approaches could also impact audiology. Future hearing-aid technology will not just restore audibility through amplification. Instead, devices will more directly alter or augment spatial aspects of the auditory scene (target identification, de-reverberation, binaural listening assistants, etc.).

Two limiting aspects of the current paper should be considered. First, the experiments presented in this paper involved natural or simulated open-field sound. Real AR applications, however, will involve fitting listeners with input transducers (presumably, microphones) that capture and analyze external audio, and output transducers (speakers or bone vibrators) the deliver the processed sound. That raises important questions about how perceptual weighting might be altered by cue distortions in the devices themselves and/or in the interaction of synthetic and natural sound. Will audio AR devices preserve the fidelity of the natural auditory scene? How are the binaural cues that listeners use affected by electroacoustic processing and reproduction? How does reverberation factor in? In our lab, we have investigated these issues through ear-canal recordings with binaural signal-processing devices in individual listeners. Diedesch [10] made such recordings made with occluding and open-fit devices across a range of simulated room conditions. Computational analyses extracted the frequency and time-specific binaural cues available in the recordings. The extracted cues can be used to power STWF analyses of stimuli as they might be experienced in auditory AR.

A second limitation of the current work is that perceptual weighting has only been assessed in brief stimuli whose components are strongly perceived to "belong" to the same auditory object. A number of studies have suggested that perceptual weighting is dramatically impacted by auditory grouping [18, 19]. This is particularly critical for VR and AR because incorrect grouping of spatial features across objects leads to binaural interference and reduced spatial awareness. It may be that the approaches described in this paper will require models of how listeners understand the auditory scene at a cognitive level. Conversely, exploiting top-down influences on perceptual organization could provide a means to dramatically augment the auditory scene, for example by spatially combining multiple distractors and isolating them from target sounds.

## 6 Summary

1. Spatial hearing by human listeners relies on auditory spatial cues that are distributed in time, frequency, and cue type.

2. Perceptual weighting of these cues strongly emphasizes robust features (e.g., rising envelope slopes) that are distributed sparsely in time and frequency.

3. Spatial cue weighting functions (e.g. TWFs) can be used to identify the spatial features most critical for compelling spatial perception. Synthesis resources can be targeted to these features.

4. This process is analogous to the use of cochlear masking patterns [7] in perceptual audio coding. It suggests the future development of "spatial codecs" that operate similarly in the spatial domain.

5. VR and AR applications will benefit from these developments by using higher-quality and more efficient spatial representations, and perceptual bases for altering and augmenting the auditory scene.

## Acknowledgments

The authors thank Asawer Nihal (Fig. 1b), Julie Stecker (Fig. 2), and Travis Moore (Fig. 3) for assistance with data collection. Portions of this work were previously described in the second author's doctoral dissertation [10] and in proceedings of the Audio Engineering Society [8] and International Congress on Acoustics [15]. Supported by NIH/NIDCD R01-DC011548 (GCS).

## References

[1] Freyman, R. L., Zurek, P. M., Balakrishnan, U., and Chiang, Y. C., "Onset dominance in lateralization," *J. Acoust. Soc. Am.*, 101(3), pp. 1649–1659, 1997.

[2] Stecker, G. C. and Hafter, E. R., "A recency effect in sound localization?" *J. Acoust. Soc. Am.*, 125(6), pp. 3914–3924, 2009, ISSN 1520-8524 (Electronic), doi:10.1121/1.3124776.

[3] Stecker, G. C., Ostreicher, J. D., and Brown, A. D., "Temporal weighting functions for interaural time and level differences. III. Temporal weighting for lateral position judgments," *J. Acoust. Soc. Am.*, 134(2), pp. 1242–52, 2013, doi: 10.1121/1.4812857.

[4] Bilsen, F. A. and Raatgever, J., "Spectral dominance in binaural lateralization," *Acustica*, 28, pp. 131–132, 1973.

[5] McFadden, D. and Pasanen, E. G., "Lateralization of high frequencies based on interaural time differences," *J Acoust Soc Am*, 59(3), pp. 634–9, 1976.

[6] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979.

[7] Fletcher, H., "Auditory patterns," *Reviews of modern physics*, 12(1), p. 47, 1940.

[8] Stecker, G. C., "Exploiting envelope fluctuations to enhance binaural perception," in *Proceedings of the Audio Engineering Society*, volume 140, 2016.

[9] Haas, H., "The influence of a single echo on the audibility of speech," *J. Audio. Eng. Soc.*, 20(2), pp. 146–159, 1972.

[10] Diedesch, A. C., *Binaural-cue weighting in sound localization with open-fit hearing aids and in simulated reverberation*, Ph.D. thesis, Vanderbilt University, Nashville, TN, 2016.

[11] Saberi, K., "Observer weighting of interaural delays in filtered impulses," *Percept. Psychophys.*, 58(7), pp. 1037–1046, 1996.

[12] Stecker, G. C. and Hafter, E. R., "Temporal weighting in sound localization," *J. Acoust. Soc. Am.*, 112(3), pp. 1046–1057, 2002.

[13] Stecker, G. C., "Temporal weighting functions for interaural time and level differences. IV. Effects of carrier frequency," *J Acoust Soc Am*, 136(6), p. 3221, 2014, doi:10.1121/1.4900827.

[14] Stecker, G. C., "Trading of interaural differences in high-rate Gabor click trains," *Hear. Res.*, 268(1-2), pp. 202–12, 2010, doi:10.1016/j.heares.2010.06.002.

[15] Stecker, G. C., "Exploiting envelope fluctuations to achieve robust extraction and intelligent integration of binaural cues," in *Proceedings of the 22nd International Conference on Acoustics*, 2016.

[16] Klein-Hennig, M., Dietz, M., Hohmann, V., and Ewert, S. D., "The influence of different segments of the ongoing envelope on sensitivity to interaural time delays," *J Acoust Soc Am*, 129(6), pp. 3856–72, 2011, doi:10.1121/1.3585847.

[17] Dietz, M., Marquardt, T., Salminen, N. H., and McAlpine, D., "Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds," *Proc Natl Acad Sci U S A*, 110(37), pp. 15151–6, 2013, doi:10.1073/pnas.1309712110.

[18] Hill, N. I. and Darwin, C. J., "Lateralization of a perturbed harmonic: effects of onset asynchrony and mistuning," *J Acoust Soc Am*, 100(4 Pt 1), pp. 2352–64, 1996.

[19] Best, V., Gallun, F. J., Carlile, S., and Shinn-Cunningham, B. G., "Binaural interference and auditory grouping," *J Acoust Soc Am*, 121(2), pp. 1070–6, 2007.