ICA 2016 BUENOS AIRES
Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century...

PROCEEDINGS of the 22nd International Congress on Acoustics

The Technology of Binaural Listening & Understanding: Paper ICA2016-445

Exploiting envelope fluctuations to achieve robust extraction and intelligent integration of binaural cues

G. Christopher Stecker[(a)]

[(a)]Vanderbilt University School of Medicine, Nashville, United States,
cstecker@spatialhearing.org

Abstract

The human auditory system achieves remarkably robust communication performance, even in complex environments featuring multiple talkers, distracting noises, echoes, and reverberation. Although the neural mechanisms of this facility are not well understood, many studies point to the importance of binaural and spatial cues present at sound onset or during other fluctuations of the temporal envelope. Specifically, transient increases in the amplitude envelope appear to trigger the sampling of binaural information, independent of binaural-cue type or frequency range. This paper begins with a review of the psychophysical and neural evidence for such a triggering process, and an exploration of signal-processing algorithms that mimic and/or exploit that process. Such algorithms can be applied in two key directions of importance to communication acoustics: First, temporal envelopes are used to guide the strategic application of spatial cues in spatial sound synthesis for human listeners. Second, temporal fluctuations are used to guide the extraction of spatial cues from binaural recordings and intelligently group those cues into temporally and spatially coherent binaural proto-objects. These applications provide critical tests of the triggering hypothesis, the general role of temporal envelope fluctuations in binaural hearing, and the neural mechanisms of integrated binaural perception. Further, they provide powerful tools for the design of efficient audio communication systems and devices that interface with human participants in real or virtual spatial settings. Supported by NIH DC011548.

Keywords: Binaural hearing, sound localization, precedence effects

Exploiting envelope fluctuations to achieve robust extraction and intelligent integration of binaural cues

1 Introduction: Envelope fluctuations enhance binaural sensitivity

Spatial hearing integrates multiple time-varying features ("spatial cues") of sounds, including interaural time and level differences (ITD and ILD) in the sound arriving at the two ears. Decades of research have described human listeners' sensitivity to these cues and the initial brain mechanisms that compute them, yet a full understanding of how the brain represents spatial information remains elusive. Current models of spatial hearing also fail to capture human listeners' ability to accurately perceive complex spatial mixtures, particularly in acoustically complex, reverberant, scenes.

The temporal envelopes of sounds strongly impact auditory perception of timbre, duration, and perceptual grouping. Envelope fluctuations are also necessary for the conveyance of ITD in high-frequency, amplitude-modulated (AM) sound, as evidenced by the dominance of sound onsets in ITD-based localization [1, 2]. Less obvious, but nonetheless critical, is the importance of envelope fluctuations for other cues, such as ILD [3, 2], or fine-structure ITD [4, 5]. The results of these studies demonstrate that envelope fluctuations such as sound onsets are necessary for processing all aspects of binaural information in periodic sounds.

In contrast to the strong onset dominance observed for periodic tones and AM sounds, studies of binaural sensitivity for noise [6, 7] and for aperiodic AM sounds [8] suggest greater binaural sensitivity during the ongoing waveform than at sound onset. Those results suggest that binaural sensitivity is gated by envelope fluctuations occurring within, rather than across, auditory critical bands. Whether those fluctuations arise intrinsically within the auditory filter (in the case of noise) or due to the overall temporal envelope (for periodic sounds), the mechanism—and its consequence—is the same.

In this paper, we review the literature on this topic and describe two brief experiments that quantify these effects and their potential applications to machine listening and spatial audio synthesis.

1.1 Onset dominance for ITD in periodically modulated high-frequency tones

In a classic observation, Hafter and Dye [1] measured ITD thresholds for trains of narrowband-filtered clicks (approximately 1/2 octave centered at 4 kHz). When clicks repeated at an interval longer than 10 ms, thresholds improved optimally with duration. Shorter interclick intervals (ICI), however, produced shallower threshold-duration functions. For ICI of 1–2 ms, thresholds for trains of 16 or 32 clicks were hardly better than for single clicks. Hafter and Dye concluded that listeners localize high-rate (short ICI) stimuli on the basis of the overall onset. At slower rates, the ongoing information contributes more. That finding has since been replicated more directly, by comparing binaural cues applied strategically to onset vs later portions of the sound [9], and by direct measurement of temporal-weighting functions (TWFs) [10, 11, 2, 12] that reveal onset dominance in the form of perceptual weighting of only the very first click—not of early clicks in

general (see Fig. 1). That finding strongly suggests that the onset *per se*—perhaps the initial rise of the amplitude envelope—drives this effect.

1.2 Onset dominance for ILD in periodically modulated high-frequency tones

Because the salience of envelope ITD can be reduced by peripheral mechanisms that smear out envelope modulations at high rates, Hafter and colleagues [3] replicated their previous experiment [1] with ILD as the discriminated cue. Other aspects of the experimental stimuli and procedure were identical. The results revealed a similar degree of onset dominance for both ITD and ILD, and a similar dependence on the click rate, leading the authors to argue against a purely peripheral account of onset dominance. There is little reason to expect ILD sensitivity to depend on envelope features, and thus no reason to expect rate-dependent onset dominance due to acoustical or peripheral effects. Rather, it appears that onset dominance for binaural discimination reflects a general characteristic of binaural neurons in the central auditory system. Subsequent studies measuring TWFs [11, 2, 12] have consistently supported this view, revealing strong ICI-dependent onset dominance for ILD similar to ITD.

1.3 Onset dominance for fine-structure ITD in low-frequency tones

A third type of binaural cue is the ITD carried by the temporal fine structure of low-frequency (e.g. 500 Hz) sounds. Because the system is sensitive to cycle-by-cycle phase differences, one expects the system to integrate information across cycles in a more-or-less optimal fashion, as would occur for cross-correlation over a reasonably long term (say, a few hundred ms). In particular, there is little reason to expect onsets or other envelope fluctuations to have any substantive effect on low-frequency ITD sensitivity. Recent studies, however, reveal a temporal dynamics for low-frequency fine-structure ITD that does not markedly differ from that described above for high-frequency envelope ITD and ILD.

First, the discrimination of fine-structure ITD fails to improve optimally with sound duration. Rather, the slope of threshold improvement with duration is nearly identical for narrow bands of noise centered at 500 Hz [13], 500-Hz pure tones [5], and click trains at 4 kHz [1]. Second, fine-structure ITD discrimination is markedly better when the cue is available at sound onset than when not [5, 14]. These observations demonstrate strong onset dominance for pure tones, even when the onset itself was diotic. Thus, it appears that even fine-structure ITD is enhanced during the early, onset, portion of the sound.

When low-frequency tones are amplitude modulated at a slow rate (32 Hz), their lateral perception is dominated by the ITD coinciding with the early rising portion of each cycle of the periodic modulator [4], rather than the portion with highest amplitude[1]. That is, the envelope fluctuations imposed by AM enhance sensitivity to ITD only within a brief window triggered by the onset of each modulation cycle. As is the case for single tone bursts [5] and for high-rate click trains [9], ITD perception is thus mediated by the early, rising, portion of each modulation cycle. It is important to note that the relatively slow 32 Hz modulation rate employed by [4] is well within the range previously shown to exhibit optimal threshold-duration slopes (i.e. equal

[1] In that study, the target was a "binaural beat" stimulus with time-varying ITD.

sensitivity to ITD in each modulation period) [1]. Thus, although the results reveal strong "onset" dominance *within* each modulation cycle, integration *across* modulation cycles is likely to be optimal, contributing to listeners' reliable perception of a single stable location.

1.4 Little to no onset dominance for "noise"

With few exceptions, the stimuli described in the previous sections were periodic: pure tones, regular AM, or click trains with constant ICI. Numerous studies have shown aperiodic stimuli, in contrast, to support greater sensitivity to *ongoing* binaural information than to *onset* cues. For example, TWFs reveal strong onset dominance for trains of repeating noise bursts, but much weaker onset dominance for trains of non-repeating noise bursts [15]. In both cases, the broadband envelopes are themselves periodic (at a rate of 500 Hz); however, within any narrow frequency band (i.e. at any single place along the basilar membrane of the inner ear), only the repeated bursts produce regular activation. Non-repeating bursts produce temporally irregular narrowband envelopes with occasional large fluctuations similar to those introduced by sound onsets or slow AM. We argue that such fluctuations explain the difference in temporal weighting of binaural information for periodic sounds and for "noise:" sensitivity to ongoing cues is driven by the aperiodic nature of the narrowband envelope, not by the broadband spectrum of the noise itself.

2 Experiment 1: Temporal weighting of binaural cues in ampitude-modulated noises

Fig. 1 illustrates the results of a brief experiment measuring TWFs for trains of amplitude-modulated noise burst trains. Stimuli comprised 16 repeating white-noise bursts, each 1 ms in duration and presented at an inter-burst interval of 2 ms (i.e., 500 Hz burst rate). ITD and ILD varied from trial to trial, with additional independent variation from burst to burst within each trial [2, 15]. Burst amplitudes were either constant or modulated 100% by a sinusoid of length one, two, or four cycles. Listeners made lateralization judgments on each trial. TWFs were estimated using multiple linear regression of rank-transformed responses onto the ITD/ILD values applied to each click. Blue lines in each lower panel of Fig. 1 plot mean (± 1 s.e.m.) beta weights across 8 listeners. Strong onset dominance was observed in the unmodulated control condition (upper left). Amplitude modulation altered this pattern, but weights were consistently highest during the rising slope of the applied modulator. Most importantly, the most intense noise bursts did not strongly influence listeners' judgments; rather the earliest (onset) burst wielded the strongest influence, even when it had the *lowest* amplitude overall. The result quantifies the importance of positive envelope fluctuations in binaural-cue processing, and identifies the temporal position that is most critical for determining perceived location.

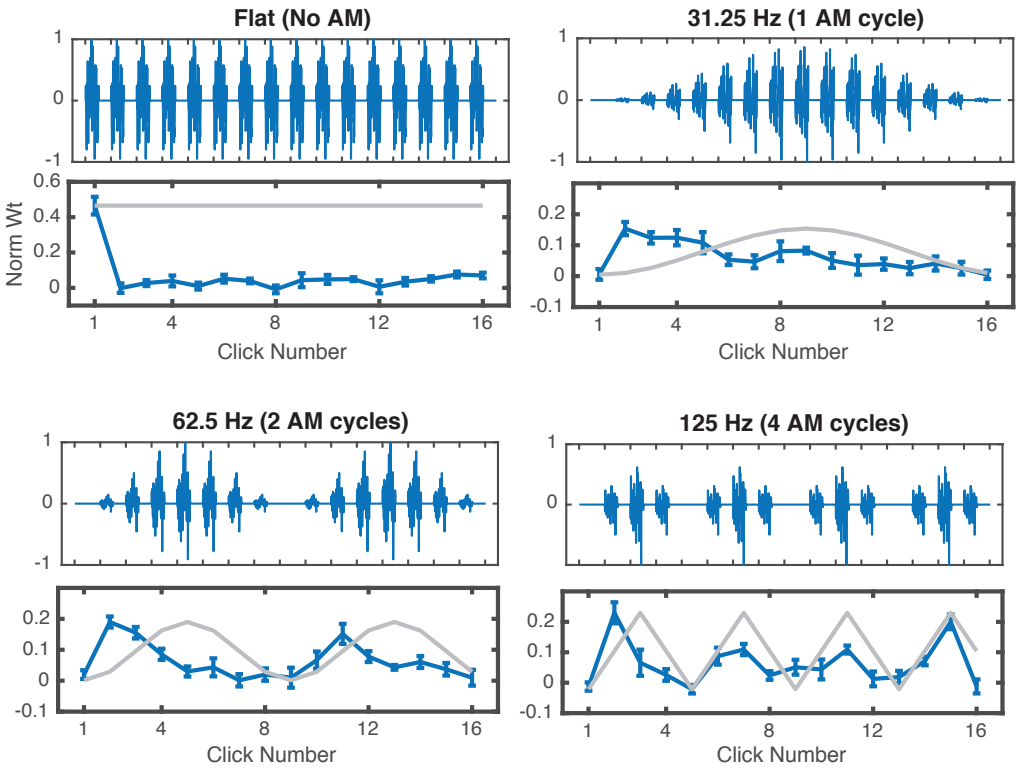

Figure 1: **Experiment 1: Temporal weighting functions (TWF) for sinudoidally amplitude-modulated trains of noise bursts.** Each pair of panels plots an example of the stimulus waveform (top) alongside the mean TWF obtained in that condition (bottom). The sinusoidal envelope applied in each condition is displayed in each lower panel (gray lines). Note that some bursts received zero amplitude. In every condition, largest weights were obtained for clicks in the early, rising, portion of the envelope.

3 Experiment 2: Exploiting envelope fluctuations for spatial sound synthesis

A second experiment exploits listeners' envelope-dependent weighting of spatial cues to alter localization judgments. As illustrated in Fig. 2, a four-channel vocoder decomposed single-syllable words into frequency bands centered at 1, 2, 4, and 8 kHz. The envelope in each band was used to modulate a continuous train of Gabor clicks centered at the corresponding analysis frequency and repeating at an ICI of 2, 5, or 10 ms. Individual clicks were directed to one of two audio channels based on the slope of the modulating envelope: rising (blue) or falling (red). A corpus of nine words was processed in this manner: "chalk," "dime," "gap," "met," "pool," "puff," "sell," "take," and "which." Listeners made localization judgments of vocoded sounds presented from a loudspeaker array (2m radius) in an anechoic chamber. Clicks corresponding to the rising and falling segments were presented from different loudspeakers separated by $\pm 11^{\circ}$,

while the overall stimulus varied from trial to trial over a range of $\pm 45°$. Stimuli were presented in anechoic conditions and in a simulated-room condition, in which each stimulus channel was convolved with a 64-channel room impulse response conveying 13 orders of reflection ($\alpha = 0.5$) in a 10m X 10m virtual room. Localization weights were calculated via multiple regression of response azimuth onto rise and fall azimuth. Bars in Fig. 2c plot the mean weights $\pm$ the full range across 3 listeners. Listeners consistently localized in the direction of the rising-envelope segments, by a factor of 2:1 to 4:1 in anechoic conditions. Consistent with the existing literature, that factor was greater at short than long ICI. When stimuli were presented in the simulated room (far left), the dominance of rising segments increased dramatically (> 9:1).

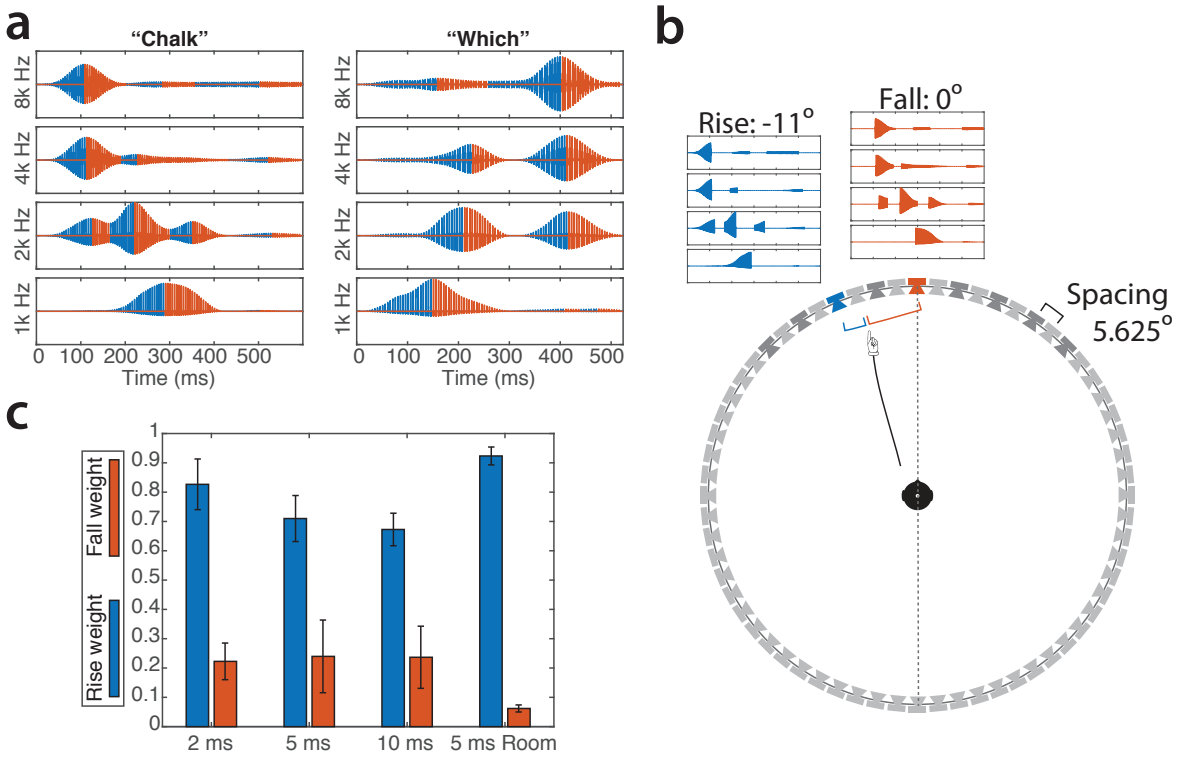

Figure 2: **Experiment 2: Spatial synthesis of vocoded speech with separated rising-envelope and falling-envelope segments.** a) Vocoded speech sounds, separated into rising (blue) or falling (red) envelope segments. b) Illustration of the loudspeaker array with rising and falling segments presented from different locations. c) Localization weights for rising (blue) and falling (red) segments. Bar heights plot the mean, and error bars plot the full range, of weights across 3 listeners. Listeners consistently localized in the direction of the rising-envelope segments, particularly at short ICI or when sounds were presented in a simulated room.

4 Conceptual model of these effects

The evidence cited above suggests a critical role for envelope fluctuations in the sampling of binaural information, regardless of the type of cue (ITD or ILD) or the frequency content of the sounds. The view that emerges from a complete consideration of the literature on this topic is that binaural cues are not processed except when "triggered" by envelope fluctuations in the form of sound onsets, slow periodic modulations, or temporally irregular fluctuations of the envelope within auditory bands.

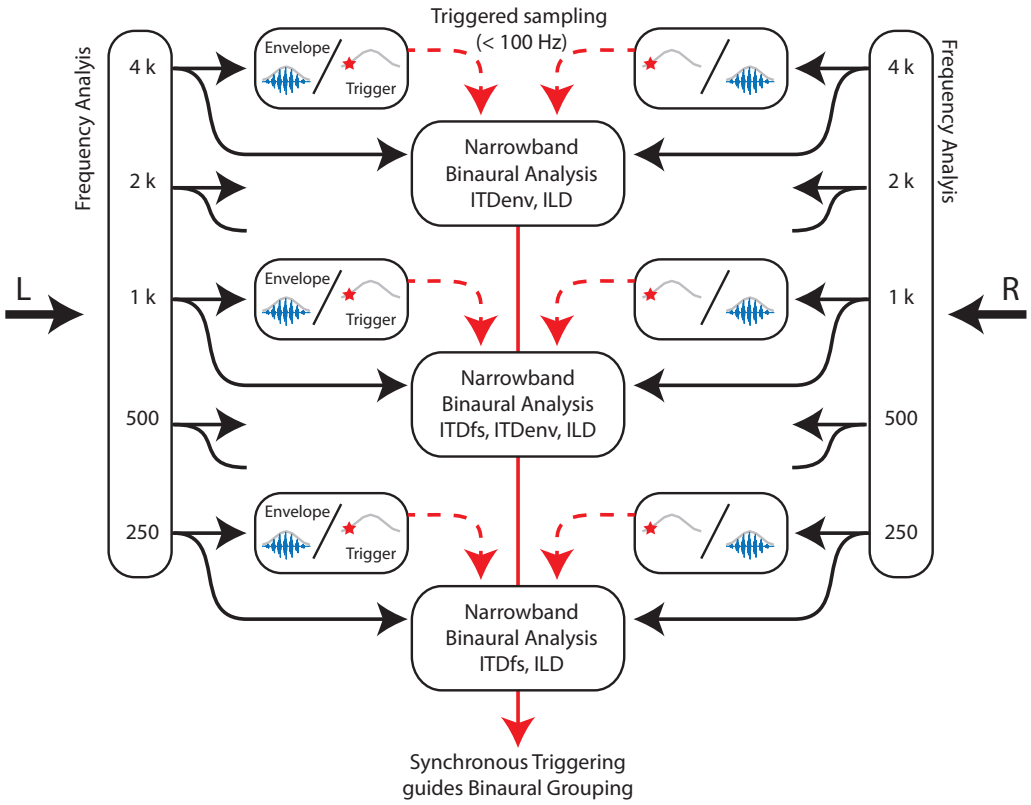

Figure 3: **Conceptual model of envelope-slope-triggered binaural sampling.** Within each frequency channel, envelope fluctuations trigger the sampling of binaural information (dashed red lines) to be made available for higher-level processing (solid red line). The triggering process is similar regardless of the frequency or type of cue (ITD/ILD) analyzed.

Fig. 3 presents a schematic model of this mechanism. In it, the output of each peripheral frequency channel is directed to binaural analysis, and to a stage of envelope processing which extracts "triggers" from the early rising segment of each fluctuation [16]. The trigger signals (dashed red lines) are used to sample binaural information in the corresponding frequency band and pass that information on to higher-level analyses (e.g., cross-frequency analysis; solid red line). Note that although an active triggering and sampling process is depicted here, the actual neural mechanism could involve different mechanisms, such as response adaptation

or delayed inhibition in the auditory brainstem [17, 18].

Several other aspects of the model should also be noted: First, each frequency favors a different mix of binaural-cue analyses—e.g. fine-structure ITD at low frequencies and envelope ITD at high frequencies—but the sampling process is similar across channels. Second, rate-dependence is captured by the triggering process, which is limited to relatively slow rates (≤ 100 Hz) due to refractoriness and/or the time constant of envelope extraction. Third, synchronous triggering events can guide the grouping of binaural information across frequency bands to perceptually define binaural "events" or proto-objects.

5 Potential applications

The mechanism described here affords human listeners with a means to sparsely sample the binaural scene and avoid many problems associated with listening in reverberation or multi-talker environments. Application of these principles to machine listening should afford many of these same advantages. For example, the components of a spatial mixture may be segregated and localized on the basis of binaural cues that coincide with infrequent fluctuations in the independent envelopes of each component.

As illustrated in Fig. 2, envelope fluctuations can also be exploited in spatial audio synthesis. The temporal envelopes of source material can be modified to control the salience of binaural information, and thus the perceived location of source objects in a synthesized scene. Alternately, binaural differences can be applied strategically to coincide with naturally occurring fluctuations. Future work in this area may lead to applications in dynamic binaural panning and in perceptually driven data compression of spatial audio.

6 Discussion

The results of these studies have clear implications for how the brain extracts spatial cues from naturally fluctuating sounds such as human speech, and how that process is altered by echoes, reverberation, and competing sources in real auditory scenes. In fact, they dramatically change our view of how the brain tracks objects in a spatial scene: rather than continuous processing of spatial information, it appears that sound envelopes form the basis for discrete and temporally sparse sampling of sound-source locations.

7 Summary

1. Binaural perception is dominated by cues that coincide with positive envelope fluctuations, such as sound onsets.

2. Onset dominance is observed consistently across binaural cue type and frequency range, affecting amplitude-modulated sounds faster than ~ 100 Hz, as well as unmodulated sounds.

3. Slowly modulated sounds (< 100 Hz AM rate), and sounds with random within-band envelope fluctuations, do not experience onset dominance.

4. Published results thus suggest that binaural information is sampled discretely, at moments of infrequent positive-going envelope fluctuations. Potential benefits of this sparse sampling could be clearer segregation of competing sounds and robust rejection of echoes and reverberation.

5. Applications to machine listening include envelope-slope-weighting of binaural information for spatial processing in reverberation and spatiotemporal grouping.

6. Applications to spatial sound synthesis utilize psychoacoustic models of dynamic binaural sensitivity to achieve robust spatial perception and data compression of spatial audio.

Acknowledgments

Anna Diedesch contributed to earlier versions of this presentation. Portions of this work were presented previously to the Acoustical Society of America (2014, Indianapolis IN), Association for Research in Otolaryngology (2015, Baltimore MD), and Audio Engineering Society (2016, Paris France). Supported by NIH/NIDCD R01DC011548.

References

[1] E. R. Hafter and R. H. Jr Dye. Detection of interaural differences of time in trains of high-frequency clicks as a function of interclick interval and number. *J. Acoust. Soc. Am.*, 73(2):644–651, 1983.

[2] G Christopher Stecker, Jennifer D Ostreicher, and Andrew D Brown. Temporal weighting functions for interaural time and level differences. III. Temporal weighting for lateral position judgments. *J. Acoust. Soc. Am.*, 134(2):1242–52, Aug 2013.

[3] E. R. Hafter, R. H. Jr Dye, and E. M. Wenzel. Detection of interaural differences of intensity in trains of high-frequency clicks as a function of interclick interval and number. *J. Acoust. Soc. Am.*, 73:1708–1713, 1983.

[4] Mathias Dietz, Torsten Marquardt, Nelli H Salminen, and David McAlpine. 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):15151–6, Sep 2013.

[5] G C Stecker and J M Bibee. Nonuniform temporal weighting of interaural time differences in 500 hz tones. *J. Acoust. Soc. Am.*, 135(6):3541–3547, 2014.

[6] J. V. Tobias and E. R. Schubert. Effective onset duration of auditory stimuli. *J. Acoust. Soc. Am.*, 31:1595–1605, 1959.

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

[8] Andrew D Brown and G Christopher Stecker. Temporal weighting functions for interaural time and level differences. II. The effect of binaurally synchronous temporal jitter. *J. Acoust. Soc. Am.*, 129(1):293–300, Jan 2011.

[9] G. C. Stecker and A. D. Brown. Temporal weighting of binaural cues revealed by detection of dynamic interaural differences in high-rate gabor click trains. *J. Acoust. Soc. Am.*, 127(5):3092–3103, 2010.

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

[11] A. D. Brown and G. C. Stecker. Temporal weighting of interaural time and level differences in high-rate click trains. *J. Acoust. Soc. Am.*, 128(1):332–341, 2010.

[12] G Christopher Stecker. Temporal weighting functions for interaural time and level differences. iv. effects of carrier frequency. *J Acoust Soc Am*, 136(6):3221, Dec 2014.

[13] T Houtgast and R Plomp. Lateralization threshold of a signal in noise. *J. Acoust. Soc. Am.*, 44(3):807–812, Sep 1968.

[14] Anna C Diedesch and G Christopher Stecker. Temporal weighting of binaural information at low frequencies: Discrimination of dynamic interaural time and level differences. *J Acoust Soc Am*, 138(1):125–33, Jul 2015.

[15] G C Stecker. Effects of carrier frequency and bandwidth on temporal weighting of binaural differences. *Proc. Meet. Acoust.*, 19:050166, 2013.

[16] P Heil. Representation of sound onsets in the auditory system. *Audiology and Neuro-Otology*, 6:167–172, 2001.

[17] E. R. Hafter, T. N. Buell, and V. M. Richards. Onset-coding in lateralization: its form site, and function. In G. M. Edelman, W. E. Gall, and W. M. Cowan, editors, *Auditory function: neurobiological bases of hearing*, pages 647–676. John Wiley & Sons, New York, 1988.

[18] R. E. Wickesberg and D. Oertel. Delayed, frequency-specific inhibition in the cochlear nuclei of mice: a mechanism for monaural echo suppression. *J. Neurosci.*, 10:1762–1768, 1990.