A new auditory theory and its implications for the study of timbre

Braden N. Maxwell[1]
Johanna B. Fritzinger[2]
Laurel H. Carney[2,3†]
[1] Eastman School of Music, University of Rochester, Rochester, NY, USA
[2]School of Medicine and Dentistry, University of Rochester, Rochester, NY, USA
[3] Hajim School of Engineering and Applied Sciences, University of Rochester, Rochester, NY, USA
[†] Corresponding author: Laurel_Carney@rochester.edu
Published 16 December 2021; https://doi.org/10.18061/FDMC.2021.0051
Author video presentation and/or other conference material: https://doi.org/10.17605/OSF.IO/XA2TZ

Abstract

A relatively new auditory theory describes how representations of the spectrum are transformed and sharpened in the early (below the cortex, or *sub-cortical*) auditory system (Carney, 2018). The current article introduces this theory and considers implications for timbre using computer model simulations. Models suggest that between two locations in the early auditory system, the auditory nerve and the midbrain, the neural representation of spectral peaks (the representation in overall activity along the tonotopic axis) becomes more precise. This peak-sharpening process depends on timing patterns of nerve activity called neural fluctuations. Neural fluctuations are comparable to temporal amplitude modulation but are, in some cases, created and modified within the auditory system rather than simply reflecting the stimulus itself. After the peak-sharpening process, a center of mass of the most prominent spectral peaks - as encoded in the midbrain - serves as a neural representation of brightness. This work suggests that brightness may be fundamentally related to the concepts of locally prominent spectral peaks (spectral irregularity) and temporal modulation.

KEYWORDS: *timbre, centroid, modulation, computational modeling, spectral irregularity*

Introduction

The spectral centroid and similar acoustic descriptors have been established as important to the perceptual representation of musical instrument sounds, and specifically the percept of brightness, in numerous previous studies (McAdams, 2019). Here we discuss how a neural representation of spectral centroid may arise in the auditory system, according to a new theory of sub-cortical auditory representations (Carney, 2018). The primary goal of our present work is not to offer an alternative calculation for the centroid (see for example various centroid definitions discussed in Marozeau et al., 2003), but rather to investigate how the representation of the spectral centroid changes throughout the early auditory system. Understanding encoding of timbre at these early stages is an important open problem, especially for the purpose of restoring the perception of timbre for individuals with hearing loss.

The theory of sound representations proposed by Carney (2018) suggests a loose analogy between the auditory midbrain and the optic nerve in the visual system. In the visual system, contrast across the visual field is more important than brightness at any given point; similarly, this auditory theory suggests that contrast in the temporal features of neural signals (properties related to their timing) across the tonotopic axis (the axis describing neural tuning, from low to high frequencies) is more important than energy at a given frequency in a sound spectrum.

A particular category of temporal features of neural signals, referred to as 'neural fluctuations,' is foundational to this theory. Neural fluctuations are relatively low-frequency (~10-200 Hz) time-varying changes in the statistics of auditory-nerve activity (Carney, 2018). This range of frequencies extends higher than the range of temporal modulation frequencies typically considered most essential for speech (Elhilali, 2019), and is generally much lower than the tonotopic 'tuning' of the neurons that carry them. These time-varying changes are often superimposed on higher-frequency changes in a neuron's activity (for example, phase locking), shaping the temporal envelope of the neural signal. Fluctuations can occur for a variety of reasons. Although they can result directly from amplitude modulation in a stimulus, they can also result from the filtering process. Beating between the components of a complex tone, for example, can produce fluctuations at the beating frequency. (Relatively high neural beating frequencies are possible at the high sound levels used for music listening (Epstein et al., 2010) because cochlear filters are wider at higher than at lower sound levels, see Figure 1 in Rose et al., 1971). Nonlinearities associated with the conversion of mechanical sound signals to neural signals strongly affect these neural fluctuations, making them substantially different from pre-neural

representations of the sound (Carney, 2018). An important property of fluctuations, owed in part to these nonlinearities, is that they are relatively consistent if the sound is rescaled to a different sound level within the music-listening range.

Most midbrain neurons are sensitive to neural fluctuations (meaning their amount of activity increases or decreases as fluctuations change). In fact, the range of frequencies we have specified for neural fluctuations is derived from midbrain properties (see Figure 9 in Joris et al., 2004). The midbrain converts differences in fluctuations across the auditory-nerve tonotopic axis (passed through several intermediate neural stages) into different amounts of activity across its own tonotopic population of neurons. Here we focus on one type of midbrain neuron, called 'band-suppressed,' that has decreased response activity when fluctuation amplitudes increase, and vice versa.

What does this theory mean for the sub-cortical encoding of spectral centroid? The first step toward answering this question is to understand how neural fluctuations can facilitate sharper representations of spectral peaks. Figure 1 walks through this peak-sharpening process as shown in model responses (models described in more detail in **Method**), beginning with a stimulus from Allen and Oxenham (2014) that has a single spectral peak (Figure 1A). Figure 1B shows the long-term average of auditory-nerve activity across a tonotopic population (the closest thing to a Fourier-analysis representation in the early auditory system) in response to the Allen and Oxenham stimulus. Each point represents the activity of a nerve fiber tuned to a unique tonotopic frequency. Note that this representation of the energy at different frequencies is broad and "blurred," and previous work has suggested that this representation is inadequate for center-of-mass estimates consistent with the percept of brightness (Maxwell et al., 2020).

Figure 1C shows activity of three auditory-nerve fibers from 1B, but zoomed to reveal detailed timing patterns (instead of the average response over time, as in 1B). For nerve fibers at points on the tonotopic axis below or above the spectral peak – in this case, 1000 Hz (left) and 1400 Hz (right), the activity is 'bumpy:' there are neural fluctuations due to beating between harmonics (4 cycles shown). However, at the peak (1200 Hz) the fluctuations are substantially reduced because the highest-magnitude harmonic dominates: there is no beating, and the envelope of the neural signal is nearly 'flat.' Importantly, this 'flatness' only occurs for nerve fibers very close to the spectral peak on the tonotopic axis.

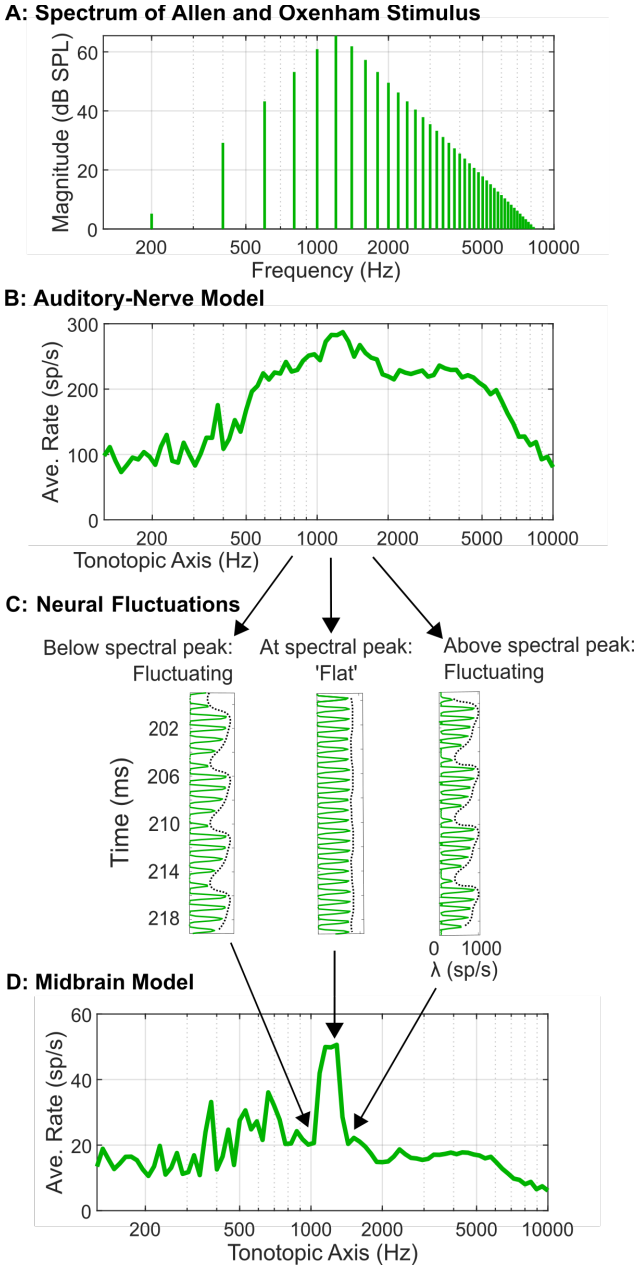

Figure 1: (A) Spectrum of a stimulus from Allen and Oxenham (2014): F0=200 Hz, spectral peak at 1200 Hz, 70 dB SPL overall level. (B) The overall amount of activity (averaged over 500 ms) along the auditory-nerve tonotopic axis in response to (A). (C) Another perspective on the activity of the same nerve fibers represented in (B), but focused on short-term timing patterns (neural fluctuations). Fluctuations are decreased at the spectral peak. For more detail on λ, see [1]. (D) Midbrain activity is sharply increased where fluctuations are decreased, and vice versa, precisely encoding the spectral peak.

After transformation by intermediate stages of the auditory system, the midbrain receives fluctuating or flat signals (Figure 1C). Midbrain band-suppressed neurons exhibit an overall decrease in activity when fluctuations increase, so they respond strongly at the spectral peak, but not above and below it (Figure 1D).

These simulations emphasize the precision of midbrain encoding of spectral features and the potential for a precise representation of brightness to emerge at the midbrain stage. In the case of a stimulus with a single spectral peak, shifting the peak (and the peak in midbrain activity) shifts the centroid. However, some spectra have multiple peaks. In results shown here, we propose that the center of mass of the peaks of the midbrain representation, regardless of how many peaks there are, serves as a sub-cortical representation of the brightness of musical sounds (a midbrain-based 'centroid'). Our results include midbrain-based centroid estimates for several instrument samples (a basic test of feasibility of this theory) and one case study, a multi-peak violin spectrum.

Method

Stimuli

For Figure 1, the stimulus was a 500-ms duration complex tone with a fundamental frequency of 200 Hz, all harmonics up to 10 kHz, a band-pass spectral envelope shape centered at 1200 Hz rolling off at 24 dB/octave, and 20-ms onset and offset ramps (Allen and Oxenham, 2014). For the simulation results shown below, stimuli were from the pre-2012 University of Iowa Musical Instrument Samples (Fritts, 1997). All instrument samples, perfomed at mezzo-forte, had a fundamental frequency of 311 Hz (E ♭ 4). Guitar and violin samples were performed on the B and D strings, respectively. Stereo recordings were converted to mono before simulations, and all stimuli were scaled to an overall level of 70 dB SPL at the input of the auditory-nerve model (this scaling is necessary because the auditory-nerve model properties change realistically with level). For simulations in which pink noise was added, the noise lasted throughout the simulation and had an approximate spectrum level of either 7 or 13 dB re 20 µPa at 100 Hz. Both levels of noise were used for Figure 2; the lower level was used in Table 1. Instrument samples were scaled to 70 dB SPL before adding noise.

Models

Computational models simulated the responses of the auditory nerve (Zilany et al., 2014) and midbrain band-suppressed neurons (Carney, et al. 2015; Maxwell et al., 2020). Eighty tonotopic frequency channels (logarithmically spaced between 125 Hz and 10 kHz) were included in each simulation. The output of each model auditory-nerve fiber was passed to brainstem and midbrain stages that inherited their input's tonotopic frequency; there was no interaction between tonotopic channels. The auditory-nerve model simulated high-spontaneous-rate fibers, the most common type of fiber (Liberman, 1978). High-spontaneous-rate fibers, with their low thresholds and saturated average response rates at high sound levels, present a challenge for traditional rate-based codes of the spectrum (Carney, 2018). Midbrain best modulation frequency (BMF, or the fluctuation frequency to which the midbrain is most sensitive) was set to 100 Hz for model midbrain neurons above 400 Hz on the tonotopic axis. Below 400 Hz, BMF was set to the tonotopic frequency divided by 4. These settings reflect a plausible range of modulation tuning frequencies for midbrain neurons (Langner et al., 2002). User-friendly software and instructions for conducting these simulations and producing similar figures is available at https://osf.io/2gpz6/. This software may be used with any input stimulus.

Neural Estimate of Centroid

To estimate the centroid as encoded in the model midbrain response (Table 1), we used a process described in detail in a previous proceedings article (Maxwell et al., 2020). Briefly, we calculated an estimate of the center of mass of the average amount of activity along the midbrain tonotopic axis (the axis shown in Figures 1D and 2C). The activity for each tonotopic location was averaged over 500 ms beginning just before the stimulus onset.

Results and Discussion

Figure 2A shows the spectrum of a violin recording with multiple local peaks (blue arrows) – two low harmonics that are higher than the others and are widely separated on a log scale (making them each a 'local peak' of sorts), and two peaks above 2000 Hz that are prominent relative to surrounding harmonics. This simulation included three versions (see **Method**), one without any added background noise (light gray), one with low-level pink noise (dark gray), and one with slightly higher-level pink noise (green, shown in 2A). As in Figure 1, the average responses of the model auditory-nerve fibers along the tonotopic axis provide a broad, non-specific

representation of the spectral peaks (Figure 2B). Figure 2C shows model midbrain activity.

In the simulation without pink noise (light gray), several harmonics are clearly represented in the midbrain, but the local spectral peaks do not stand out clearly from the other harmonics in every case. Because the violin fundamental frequency is higher than the fundamental frequency in Figure 1 (311 instead of 200 Hz) there is more empty space in the spectrum – the wider spacing of harmonics prevents beating (fluctuation) frequencies that are constructive for the peak-sharpening process. However, in many realistic listening situations this space is filled by the harmonics of other instruments or low-level ambient noise. When a low-level pink noise is introduced (2C; dark gray line) and then a higher-level pink noise (2C; green line) the spectral peak code becomes sharper, emphasizing the four peaks identified by blue arrows in the spectrum. Counterintuitively, a small amount of noise actually *enhances* this representation by increasing the amount of neural fluctuations in the auditory-nerve responses at tonotopic locations below and above the spectral peaks, while leaving fluctuations in nerve fibers near the spectral peaks unchanged (flat).

If we calculate a center of mass of this model midbrain representation (the green line in Figure 2C), the result is a 'centroid' in which locally prominent spectral peaks play a more influential role than individual harmonics. Moreover, the absolute magnitude of harmonics near the peaks matters less than their local prominence; note that the peak in the midbrain graph near 4000 Hz on the tonotopic axis is just as high as the peak near 600 Hz (Figure 2C), even though the harmonic near 600 Hz has a higher magnitude (Figure 2A). In this way, our proposed encoding of brightness is connected to the notion of spectral jaggedness or irregularity (Caclin et al., 2005).

This theoretical account of a sub-cortical neural representation of brightness suggests interesting relationships between brightness, spectral irregularity, and temporal modulations underlying the fluctuations that drive the peak-sharpening process. An important question is whether this representation can actually account for human perceptual results. We have shown previously that this simulated neural representation can potentially explain some brightness results in a psychophysical study involving single spectral peaks (Maxwell et al., 2020; Allen and Oxenham, 2014), but can the midbrain representation account for the frequently observed brightness-based perceptual organization of musical instrument timbres?

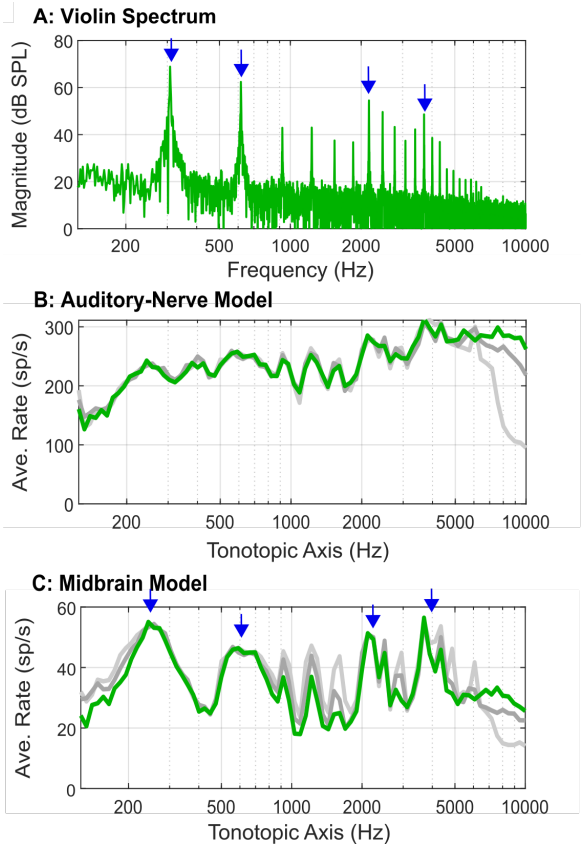

Figure 2: (A) Spectrum of violin stimulus (70 dB SPL) during steady state, with the higher-level pink noise added (see Method). (B) Auditory-nerve activity in response to (A), averaged over 500 ms, with no noise (light gray), low-level noise (dark gray), and the higher-level noise (green). (C) Elevated midbrain activity encodes prominent spectral peaks (blue arrows). Representation of spectral peaks is sharper with added noise (green, dark gray) than without noise (light gray).

The results in Table 1 offer a preliminary test of this question. 'Centroids' were estimated for simulations of midbrain responses to several different instrument samples. As in Figure 2 (dark gray lines), a low-level pink noise was added to these simulations to reflect realistic listening environments. Generally, these results are consistent with previous studies of perceptual timbre spaces: trumpet, oboe, and violin have higher values; vibraphone, clarinet, and horn have lower values; and piano and guitar are in between (compare to brightness axis in Figure 2.4 of McAdams 2019). Note that in this preliminary test, the stimuli were not directly matched to samples used in a previous perceptual study, and brightness can vary substantially for a given instrument depending on details of the sample itself (e.g.

Schoonderwaldt, 2009). Furthermore, modeling these general trends is not a very specific test of this theory of brightness encoding; as such, these results constitute a baseline test of feasibility rather than strong evidence for this specific explanation of neural timbre encoding.

*Table 1: Midbrain-model centroid estimates for selected instrument samples from the University of Iowa MIS, ordered from lowest centroid estimate to highest. *[2]*

Instrument Sample	Midbrain-Model Centroid Estimate (Hz)
Vibraphone	330
Horn	521
Clarinet	522
Bassoon	581
Trombone	701
Piano	721
Oboe	913
Trumpet	1186
Violin	1202
*Guitar	*530 (for 500 ms) *1092 (for 200 ms)

Conclusion

The neural fluctuation theory of Carney (2018) offers a novel explanation of how timbral brightness may be encoded through the sharpening of spectral-peak representations in the sub-cortical auditory system. This theoretical account suggests potentially fundamental neural relationships between brightness, temporal modulation, and spectral irregularity. Given the substantial role that time-varying changes in neural signals play in this theory, future work may examine implications for spectral flux and attack time as well.

Acknowledgements

Supported by NIH-DC010813 & NIH-DC001641.

End Notes

[1] λ can be understood as the instantaneous likelihood of an auditory-nerve spike (a single action potential), expressed here in units of spikes per second (sp/s). If the value of λ was 100 sp/s for a full second, 100 spikes would occur, on average, during that second.

[2] Estimated guitar centroid varied substantially depending on how much decay was included. Both timespans began at the same point before note onset.

References

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *JASA*, *135*(3), 1371-1379. https://doi.org/10.1121/1.4863269

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *JASA*, *118*(1), 471-482. https://doi.org/10.1121/1.1929229

Carney, L. H. (2018). Supra-threshold hearing and fluctuation profiles: Implications for sensorineural and hidden hearing loss. *Journal of the Association for Research in Otolaryngology*, *19*(4), 331-352. https://doi.org/10.1007/s10162-018-0669-5

Carney, L. H., Li, T., and McDonough, J. M. (2015). Speech coding in the brain: Representation of vowel formants by midbrain neurons tuned to sound fluctuations. *Eneuro*. *2*(4). https://doi.org/10.1523/ENEURO.0004-15.2015

Elhilali, M. (2019). Modulation representations for speech and music. In K. Siedenburg, C. Saitis, S. McAdams, A. Popper, R. Fay. (Eds.), *Timbre: Acoustics, perception, and cognition* (pp. 335-359). Springer. https://doi.org/10.1007/978-3-030-14832-4_12

Epstein, M., Marozeau, J., & Cleveland, S. (2010). Listening habits of iPod users. *Journal of Speech, Language, and Hearing Research*, *53*, 1472-1477. https://doi.org/10.1044/1092-4388(2010/09-0059)

Fritts, L. (1997). University of Iowa musical instrument samples database. Retrieved January 14, 2021 from theremin.music.uiowa.edu/MIS.html

Joris, P. X., Schreiner, C. E., and Rees, A. (2004). Neural processing of amplitude-modulated sounds. *Physiological Reviews,* 84, 541-577. https://doi.org/10.1152/physrev.00029.2003

Langner, G., Albert, M., & Briede, T. (2002). Temporal and spatial coding of periodicity information in the inferior colliculus of awake chinchilla (Chinchilla laniger). *Hearing Research*, *168*(1-2), 110-130. https://doi.org/10.1016/S0378-5955(02)00367-2

Liberman, M. C. (1978). Auditory-nerve response from cats raised in a low-noise chamber. *JASA*, *63*(2), 442-455. https://doi.org/10.1121/1.381736

Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *JASA*, *114*(5), 2946-2957. https://doi.org/10.1121/1.1618239

Maxwell, B. N., Fritzinger, J. B., Carney, L. H. (2020). Neural mechanisms for timbre: spectral-centroid discrimination based on a model of midbrain neurons. timbre2020.mus.auth.gr/assets/papers/22.Maxwell.pdf

McAdams, S. (2019). The perceptual representation of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. Popper, R. Fay. (Eds.), *Timbre: Acoustics, perception, and cognition* (pp. 23-57). Springer. https://doi.org/10.1007/978-3-030-14832-4_2

Rose, J. E., Hind, J. E., Anderson, D. J., & Brugge, J. F. (1971). Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey. *Journal of*

Neurophysiology, *34*(4), 685-699. https://doi.org/10.1152/jn.1971.34.4.685

Schoonderwaldt, E. (2009). The violinist's sound palette: spectral centroid, pitch flattening and anomalous low frequencies. *Acta Acustica United with Acustica*, *95*(5), 901-914. https://doi.org/10.3813/AAA.918221

Zilany, M. S., Bruce, I. C., & Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *JASA*, *135*(1), 283-286. https://doi.org/10.1121/1.4837815