

Spatio-temporal representation of the pitch of complex tones in the auditory nerve

Leonardo Cedolin^{1,2} and Bertrand Delgutte^{1,2,3}

¹ Eaton-Peabody Laboratory, Massachusetts Eye and Ear Infirmary,
bertrand_delgutte@meei.harvard.edu

² Speech and Hearing Bioscience and Technology Program, Harvard-M.I.T. Division of
Health Sciences and Technology, leonardo.cedolin@alum.mit.edu

³ Research Laboratory of Electronics, M.I.T.

1 Introduction

Although pitch is a fundamental auditory percept that plays an important role in music, speech, and auditory scene analysis, the neural codes and mechanisms for pitch perception are still poorly understood. In a previous study (Cedolin and Delgutte 2005), we tested the effectiveness of two classic representations for the pitch of harmonic complex tones at the level of the auditory nerve (AN) in cat: a rate-place representation based on resolved harmonics and a temporal representation based on pooled interspike-interval distributions (a.k.a. autocorrelation). Both representations supported precise pitch estimation in the F0 range of cat vocalizations (500-1000 Hz), but neither was entirely consistent with human psychophysical data. Specifically, the rate-place representation failed to predict the existence of an upper limit for the pitch of missing-F0 complex tones (Moore 1973). The rate-place representation also degrades rapidly with increasing sound level, in contrast to the relatively robust pitch discrimination performance. The interval representation did not account for the greater salience of pitch based on resolved harmonics compared to pitch based on unresolved harmonics (Carlyon and Shackleton 1994).

Here, we investigate an alternative, “*spatio-temporal*” neural representation of pitch, which may combine the advantages and overcome the limitations of the traditional rate-place and interval representations.

1.1 Spatio-temporal representation of pitch

Physiological and modeling studies have shown that the phase of basilar membrane motion in response to a pure tone varies rapidly with cochlear place near the place tuned to the tone frequency (Pfeiffer and Kim 1975). At frequencies within the range of phase-locking, this rapid spatial change in phase is reflected in the timing of AN spike discharges, thus generating a spatio-temporal cue to the frequency of

the pure tone which can in principle be extracted by a neural mechanism sensitive to the relative timing of spikes from adjacent cochlear locations (Shamma 1985).

For harmonic complex tones, such rapid changes in phase are expected to occur at each of the spatial locations tuned to a resolved harmonic, thereby providing “*spatio-temporal*” cues to pitch that could serve as input to a harmonic template mechanism. Figure 1 shows the response of a peripheral auditory model (Zhang et al., 2001) to a missing-F0 harmonic complex tone. The latency of the resulting traveling wave varies more rapidly for CFs near low-order harmonics than for CFs in between two harmonics (white broken line in Fig. 1). To extract these spatio-temporal cues, we compute the spatial derivative of the response pattern (a point-by-point difference between adjacent rows in Fig. 1), then integrate the absolute value of the derivative over time. This “*mean absolute spatial derivative*” (MASD) simulates a lateral inhibitory mechanism operating upon the spatio-temporal pattern of AN activity (Shamma 1985). The MASD shows local maxima at CFs corresponding to the frequencies of Harmonics 2-6, while the average discharge rate (R_{avg}), obtained by integrating the response at each CF over time, is largely saturated at this stimulus level. Thus, these model results suggest that spatio-temporal pitch cues may persist at levels at which the rate-place representation fails due to the saturation of AN fibers responses. A major goal of the present study is to test this prediction physiologically.

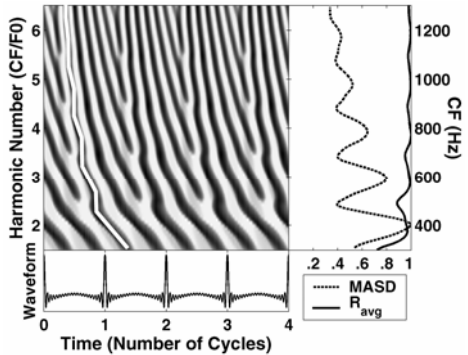


Fig. 1. Spatio-temporal response of peripheral auditory model (Zhang et al. 2001) to a harmonic complex tone with missing F0 (200 Hz). Filter bandwidths were set to match human psychophysical masking data (Moore and Glasberg 1990).

1.3 Scaling invariance in cochlear mechanics

Measuring the entire spatio-temporal response pattern of the AN for a complex-tone stimulus as in Fig. 1 would be extremely difficult, because it would require a very fine, regular and extensive sampling of CFs. We overcame this hurdle by applying the principle of local *scaling invariance* in cochlear mechanics (Zweig 1976). Scaling invariance implies that the spatio-temporal response pattern to a complex tone *with a given F0* can be inferred from the responses *at a single cochlear place (fixed CF)* to a series of complex tones with varying F0, if cochlear place and time are expressed in dimensionless units $CF/F0$ (*harmonic number*) and $t \times F0$ (*normalized time*), respectively. Figure 2 shows the model spatio-temporal response pattern to a complex tone *with fixed F0* (left) next to the model response pattern *at a fixed CF* to a series of complex tones with *varying F0* (right). The

harmonic number CF/F_0 varies from 1.5 to 4.5 in both cases. The two response patterns are nearly undistinguishable and both R_{avg} and MASD, computed as in Fig. 1, show nearly identical features, thus justifying our method for inferring the spatio-temporal responses from responses of a single AN fibers.

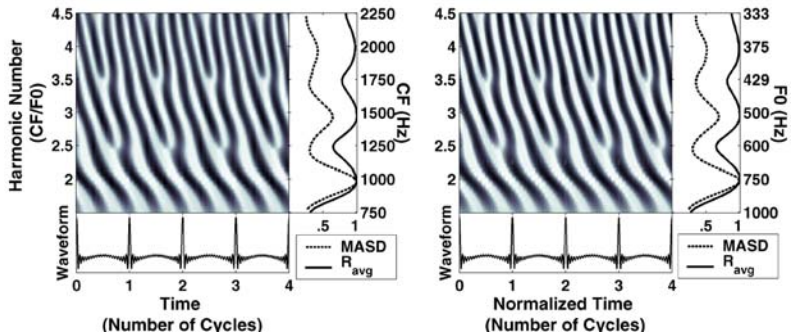


Fig. 2. Left: Spatio-temporal response pattern of Zhang et al. (2001) model to a harmonic complex tone (missing $F_0 = 500$ Hz). Right: Model responses for a single cochlear place ($CF = 1500$ Hz) to a series of complex tones with varying F_0 (333-1000 Hz). Note the normalized time scale and the inverted F_0 scale. Filter bandwidths were set to match physiological data from cat AN (Carney and Yin 1988).

2 Methods

Methods for recording from auditory-nerve fibers in anesthetized cats were as described by Cedolin and Delgutte (2005).

Stimuli were harmonic complex tones with missing F_0 s. For each fiber, the F_0 range was chosen such that the harmonic number CF/F_0 varied from 1.5 to 4.5 in order to capture low-order harmonics likely to be resolved. Each complex tone was composed of Harmonics 2 to 20, all of equal amplitude and in cosine phase. Each of the F_0 steps lasted 200 ms and was presented 40 times. The sound pressure level of each harmonic was initially set at 10-15 dB above rate threshold for a pure tone at CF . When possible, the stimulus level was then systematically varied over a 20-30 dB range.

Period histograms were constructed in response to each complex tone and displayed as a function of both normalized time and CF/F_0 for each fiber. R_{avg} and MASD were computed from the resulting response pattern as for Fig. 1.

3 Results

Figure 3 shows the responses to complex tones for three AN fibers with different CF s. For the low- CF fiber (700 Hz, A), the response latency varies very uniformly with harmonic number, indicating that the cochlear frequency selectivity is insuffi-

cient to resolve harmonics at this CF; as a result neither rate-place nor spatio-temporal pitch cues can be detected when examining R_{avg} and MASD.

The spatio-temporal response pattern of the fiber with an intermediate CF (2150 Hz, B) shows non-uniform variations in response latency with harmonic number qualitatively similar to those predicted by the model of Fig. 2. The latency varies rapidly at integer harmonic numbers, while it changes more slowly between integers. As a result, the MASD shows local maxima at Harmonics 2, 3 and 4, thus providing evidence for spatio-temporal cues to pitch in the AN response. R_{avg} also shows peaks at integer harmonic numbers, although they are less pronounced than for the MASD.

For the high-CF fiber (4.3 kHz, C), the spatio-temporal response pattern shows no evidence of phase locking to individual harmonics. As a result, the MASD is basically flat. In contrast, R_{avg} shows pronounced peaks at Harmonics 2-6, consistent with the improvement in relative frequency selectivity at higher CFs (Cedolin and Delgutte, 2005). Thus, because the spatio-temporal representation depends on phase locking, it predicts an upper F_0 -limit to pitch which is consistent with psychophysical observations but is not seen in the rate-place representation.

We have hypothesized that the spatio-temporal representation of pitch may remain effective at stimulus levels where the rate-place representation breaks down. Figure 4 shows results for one AN fiber at 3 different levels (10, 25 and 40 dB re. threshold). At the low level (A), Harmonics 2, 3, and possibly 4 appear as distinct peaks in both R_{avg} and MASD. At the intermediate level (B), R_{avg} begins to show signs of saturation as only Harmonic 2 is apparent. In contrast, strong latency cues to Harmonics 2, 3, and possibly 4 are still present in the spatio-temporal response pattern, resulting in corresponding prominent peaks in MASD. At the highest level

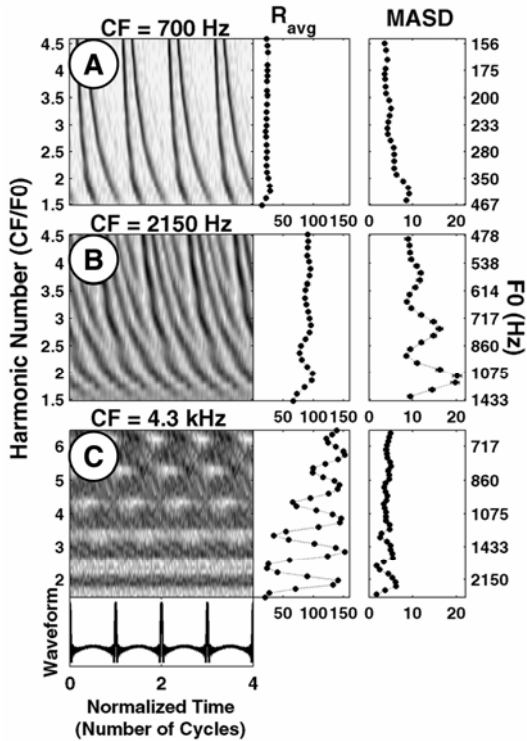


Fig. 3. Responses of 3 AN fibers with different CFs to series of harmonic complex tones. The stimuli were at 10, 20 and 30 dB, respectively, above each fiber's threshold.

(C), R_{avg} is completely saturated, while peaks at Harmonics 2 and 3 are still easily detectable in the MASD.

This example supports our hypothesis that the spatio-temporal representation is more robust with respect to stimulus level than the rate-place representation.

Intuitively, the more pronounced the oscillations in R_{avg} and MASD, the better individual harmonics are resolved and therefore the stronger the pitch cues. To quantify this intuition, we fit a damped sinusoidal function of harmonic number separately to R_{avg} and MASD, then use the area between the top and bottom envelopes of the fitted curve as a measure of the strength of the pitch representation. Since R_{avg} and MASD have different units, we express this area relative to the typical standard deviation of the data points, making this metric analogous to a d' . We call this metric the *harmonic strength* of the MASD or the R_{avg} .

To compare the strengths of the two pitch representations, we define a metric called “*normalized strength difference*” (NSD) as the difference between the harmonic strength of the MASD and that of the R_{avg} , divided by their sum. NSDs take values between -1 and 1 , with positive values indicating that MASD provides stronger pitch cues than R_{avg} .

Figure 5 shows normalized strength differences against CF for 3 different level ranges for our data set. For CFs between 1 and 3 kHz, the strengths of the two representations are comparable at low levels (A). In this CF range, the NSDs tend to be positive at moderate levels (B), indicating that MASD better represents resolved harmonics than R_{avg} . This tendency is even more pronounced at high levels (C), where virtually all NSDs are positive. Thus, the spatio-temporal representation is more robust at higher stimulus levels than the rate-place representation in this CF range.

For CFs above 3 kHz, NSDs tend to take large negative values at all levels, indicating that rate cues to resolved harmonics are stronger than latency cues. This result is most likely caused by the degradation of phase-locking with increasing frequency of the resolved harmonics near the CF. Overall, the spatio-temporal pitch representation is most effective for CFs between 1 and 3 kHz.

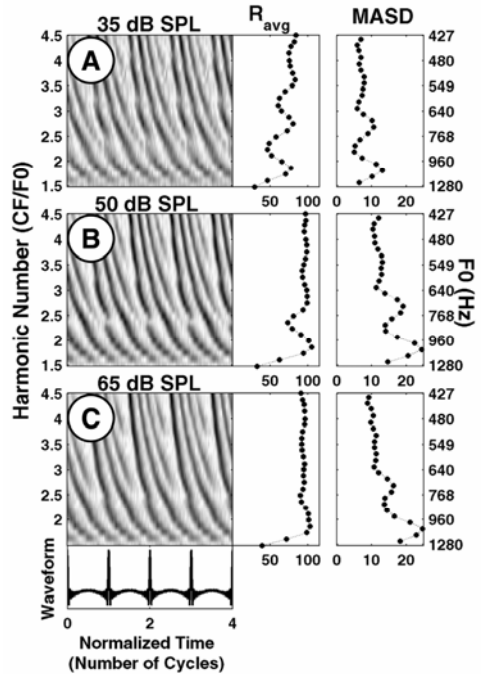


Fig. 4. Spatio-temporal response pattern of AN fiber (CF=1920 Hz) to a series of harmonic complex tones at 3 different stimulus levels. Pure-tone threshold at CF was 25 dB SPL.

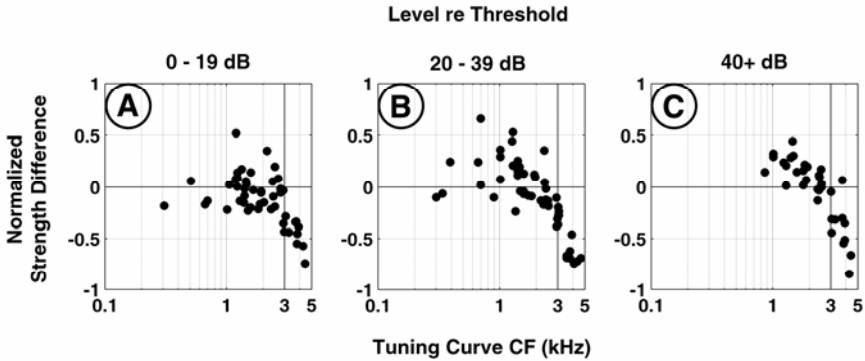


Fig. 5. Scatter plots of normalized strength difference (NSD) of MASD re. R_{avg} against CF for 3 different level ranges (expressed in dB re. pure-tone threshold at CF). Each point shows data for one AN fiber. Plots only include measurements for which MASD and R_{avg} oscillated sufficiently against harmonic number to reliably fit a damped cosine curve.

4 Discussion

We found that robust spatio-temporal cues to resolved harmonics are available in the response of AN fibers whose CFs are high enough for harmonics to be sufficiently resolved (above ~ 1 kHz in cat), but below the limit (~ 3 kHz) above which phase-locking is significantly degraded.

To translate the CF range where the spatio-temporal representation is effective into a range of stimulus F0s, we rely again on the scaling invariance principle. Since we almost always selected the F0 range for each fiber so that the harmonic number CF/F_0 varied from 1.5 to 4.5, the fiber CF was on average 3 times greater than the F0. Hence, the 1-3 kHz CF range in which the spatio-temporal representation works best approximately corresponds to an F0 range from 300 Hz to 1 kHz, which covers the entire range of cat vocalizations. These limits are approximate as they depend on the signal-to-noise ratio of our measurements from single fibers.

What might be the corresponding F0 range in humans? Because the upper limit is determined by neural phase locking, and there are no strong reasons to assume that the frequency dependence of phase-locking greatly differs among mammalian species, it may be similar in humans. On the other hand if, as argued by Shera et al (2002) (but see Ruggero et al. 2005), cochlear frequency selectivity is 2-3 times sharper in humans than in cats, the 300-Hz lower F0-limit in cats might translate to about 100 Hz in humans. If so, the F0 range where the spatio-temporal representation works best in humans would encompass most of the range of human voice.

The proposed spatio-temporal representation of pitch seems to overcome some of the main limitations of the traditional rate-place and interspike-interval representations in accounting for the main trends in human psychophysics. Unlike the rate-place representation, it predicts the existence of an upper F0 limit to the perception of the pitch of missing-F0 complex tones, and it remains effective at high stimulus

levels; unlike the interspike-interval representation, its strength depends strongly on harmonic resolvability, and it does not require long neural delays or precise neural oscillators for which there is little physiological evidence.

A key question is whether the spatio-temporal pitch cues available in the AN are extracted in the central nervous system. In principle, the spatio-temporal cues could be extracted by a neural mechanism that (1) receives inputs from auditory nerve fibers with neighboring CFs and (2) is sensitive to differences in the timing of these inputs. Two such mechanisms are lateral inhibition (Shamma, 1985) and cross-frequency coincidence detection (Carney 1990), which would likely produce *local maxima* and *local minima*, respectively, in discharge rate at locations corresponding to the harmonics, where the rapid phase changes would cause inputs with neighboring CFs to be less coincident. Since evidence for both mechanisms exists in the cochlear nucleus, this site seems to be a promising focus for future studies.

Acknowledgments

This work was supported by NIH grants DC 02258 and 05209.

References

- Carlyon RP and Shackleton TM (1994) Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? *J Acoust Soc Am* 95.
- Carney LH and Yin TCT (1988) Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. *J Neurophysiol* 60.
- Carney LH (1990) Sensitivities of cells in the anteroventral cochlear nucleus of cat to spatio-temporal discharge patterns across primary afferents. *J Neurophysiol* 64.
- Cedolin L and Delgutte B (2005) Pitch of Complex Tones: Rate-Place and Interspike Interval Representations in the Auditory Nerve. *J Neurophysiol* 94: 347–362.
- Glasberg BR and Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. *Hear Res* 47: 103-138.
- Moore BCJ (1973) Some experiments relating to the perception of complex tones. *Q J Exp Psychol* 25: 451-475.
- Pfeiffer RR and Kim DO (1975) Cochlear nerve fiber responses: Distribution along the cochlear partition. *J Acoust Soc Am* 58: 867-965.
- Shera CA, Guinan JJ, Jr., and Oxenham AJ (2002) Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proc Natl Acad Sci USA* 99.
- Shamma SA (1985) Speech processing in the auditory system. I: The representation of speech sounds in the responses of the auditory nerve. *J Acoust Soc Am* 78: 1612-1621.
- Ruggero M and Temchin AN (2005) Unexceptional sharpness of frequency tuning in the human cochlea. *Proc Natl Acad Sci USA* 102: 18614-18619.
- Zhang X, Heinz MG, Bruce IC, and Carney LH (2001) A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *J Acoust Soc Am* 109: 648-670.
- Zweig G (1976) Basilar membrane motion. *Cold Spr Harb Symp Quant Biol* 40: 619-633.