Predicting discrimination of formant frequencies in vowels with a computational model of the auditory midbrain

Laurel H. Carney and Joyce M. McDonough

***Abstract*— Neural information for encoding and processing temporal information in speech sounds occurs over different time-courses. We are interested in temporal mechanisms for neural coding of both pitch and formant frequencies of voiced sounds such as vowels. In particular, in this study we will describe a strategy for quantifying the ability to discriminate changes in spectral peaks, or formant frequencies, based on the responses of neural models. Previous studies have explored this question based on responses of computational models for the auditory periphery, that is, responses of the population of auditory-nerve (AN) fibers (e.g. [1]-[2]). In this study we quantify formant-frequency discrimination based on the responses of models for auditory midbrain neurons at the level of the inferior colliculus (IC). These neurons are tuned to both audio frequency and to low-frequency amplitude modulations, such as those associated with pitch.**

***Index Terms*— Auditory midbrain, computational neuroscience, neural coding, statistical decision theory.**

I. Introduction

Studies of temporal mechanisms for neural processing of speech have traditionally focused on phase-locking (or synchronization) of neural discharges to the stimulus fine-structure as a mechanism for coding spectral features (reviewed in [3]). A beneficial feature of phase-locking as a coding mechanism is that it is robust across a wide range of sound levels and it is also robust in noise. AN fibers are each tuned to a narrow band of frequencies, and their discharges synchronize to the fine-structure of the stimulus frequencies in that band (Fig. 1). However, AN discharges simultaneously synchronize to large, relatively slow fluctuations of the stimulus envelope (Fig. 1). For voiced sounds, these fluctuations are associated with the pitch period. This feature of the AN responses is of interest because the majority of

Manuscript received February 24, 2012. This work was supported in part by NIH grant R01-001641.

L. H. Carney is with the Departments of Biomedical Engineering and Neurobiology & Anatomy, University of Rochester, Rochester, NY 14642 USA (585-276-3948; e-mail: laurel.carney@rochester.edu).

J. M. McDonough is with the Department of Linguistics, University of Rochester, Rochester, NY 14642 USA (e-mail: joyce.mcdonough@rochester.edu).

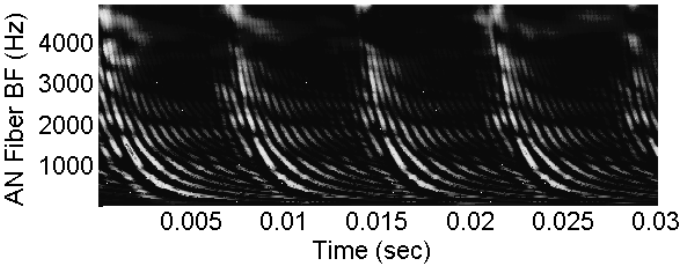


Fig. 1. Response of a population of model AN fibers tuned to frequencies from 300 to 5000 Hz to the vowel /a/. The detailed timing of responses to the fine structure of the stimulus components near formants is apparent, as well as the more global phase-locking to the pitch period, which stretches across the entire population of response fibers. The Zilany *et al.* AN model [4] was used to compute these responses.

midbrain neurons are tuned to sounds with low-frequency amplitude modulations that have modulation frequencies in the range of voice pitch.

The relative strength of phase-locking to the pitch period vs. phase-locking to higher frequency harmonics varies in an interesting manner across the AN population. In each AN fiber's response, the dominance of the phase-locking to the pitch period depends upon the relative magnitudes of the frequency components that fall within that fiber's frequency range (or bandwidth). For fibers tuned near a spectral peak, the responses to harmonics near the spectral peak are relatively sustained throughout each pitch period (Fig. 1), and the energy in the response that is phase-locked to the pitch-related periodicity is relatively weak (Fig. 2A, left). For fibers tuned to frequency channels away from spectral peaks, in which the spectral components are similar in amplitude, responses are strongly periodic at the fundamental frequency (Fig. 1 and Fig. 2A, right). In these frequency channels, AN responses to harmonics with similar amplitudes result in "beats" at the frequency difference of the components; for voiced speech sounds, the difference frequency is the fundamental frequency (F0) which is the voice pitch.

Many auditory neurons in the midbrain (inferior colliculus, IC) and cortex are tuned to low-frequency amplitude modulations or periodicities ([5]-[7]). Each IC cell has a best audio frequency that is inherited from the tuning of its neural inputs and a best modulation frequency that arises at the level of the IC itself, presumably due to neural circuitry, such as

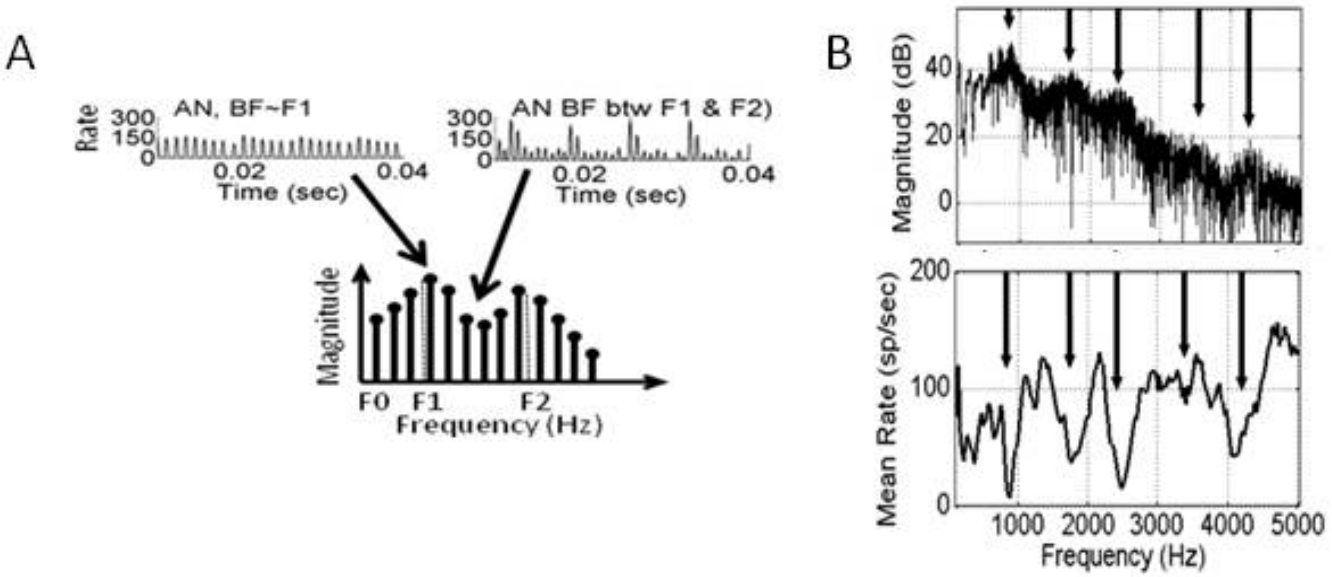


Fig. 2. A) Illustration of two model AN fiber responses to a vowel sound. One fiber (left) is tuned to a frequency near a spectral peak, resulting in a response that is dominated by the frequency associated with that peak. The other fiber (right) is tuned to a frequency between formant peaks; this fiber's response has a strong component that is phase-locked to the pitch, which is the frequency difference between the frequency components to which this fiber responds. B) The spectrum of the vowel */a/* (top) and responses of model midbrain responses (bottom). Decreases in average rate occur for model neurons tuned near formant peaks in the speech stimulus. The AN responses were simulated using the Zilany *et al.* AN model [4]. Midbrain responses were computed using the model of Nelson and Carney [8].

interactions between inhibitory and excitatory inputs. Thus, a midbrain neuron may receive inputs that are tuned to a best audio frequency of 3 kHz, but it will respond best when those inputs are temporal modulated at a particular low-frequency (e.g. 100 Hz, as in Fig. 2A, right). Several computational models have been proposed for AM tuning at the level of the IC (e.g. [8]-[12]; see [13] for a review). The low-frequency periodicity, or temporal modulation frequency, that elicits the best response in a midbrain cell is referred to as its best modulation frequency (BMF). The majority of tuned IC cells have BMFs in the voice pitch range [7]. Voiced sounds elicit strong periodicities in many frequency channels, with the degree of modulation varying depending upon proximity to formants (Fig. 2A). The pitch of a voiced sound determines the subset of midbrain neurons that respond most strongly, and fluctuations in pitch over time will result in dynamic shifts in the response across the population of these neurons.

The *strength of temporal fluctuations* within a narrow audio frequency band is the essential stimulus for many central auditory neurons (as opposed to just the presence of energy within the frequency band.) The brain apparently parses sound into a two-dimensional representation (at least), with *best audio frequency* being one frequency dimension, and *best modulation frequency* another. There's some evidence that both of these frequency axes are represented topographically in the brain, in orthogonal dimensions [14].

Changes in discharge rate across the group of central neurons that respond to a given voiced sound encode the frequencies of formant peaks. As illustrated above (Figs. 1, 2A), frequency channels near formants have responses that are more weakly modulated at the fundamental frequency than frequency channels away from formants. Thus, periodicity-tuned midbrain neurons with BFs *near* formants will have *weaker* responses than midbrain neurons with BFs between formants. These response properties of midbrain neurons suggest a counter-intuitive *drop* in rate for midbrain cells tuned near formant frequencies (Fig. 2B). This prediction, based on neural model responses, is consistent with preliminary physiological recordings (not shown). The prediction is also consistent with the established phenomenon of "locking suppression" that has been illustrated in central auditory neurons with stimuli that included narrowband peaks in the context of a wideband background [15].

II. Predicting the Ability to Discriminate Formant Frequencies

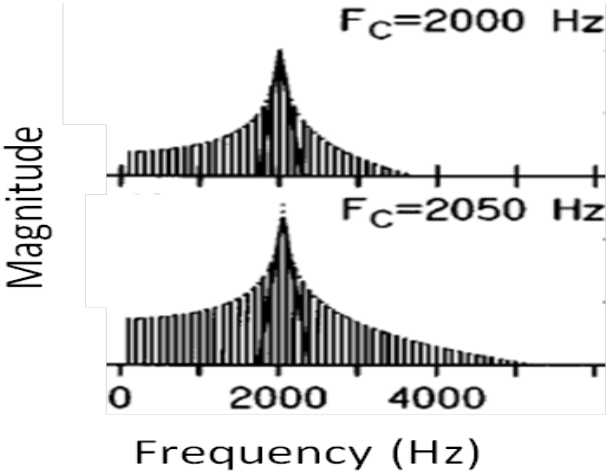
In order to better understand neural mechanisms for processing temporal aspects of speech, we must understand how the brain responds to not only the energy vs. frequency (i.e. classical spectral energy), but also to *temporal fluctuations* in energy within each frequency channel. We are exploring how these temporal fluctuations vary with spectral features and how they interact with the two-dimensional frequency tuning of auditory neurons, in which each neuron is characterized by both its best audio frequency and by its best modulation frequency.

A. Stimuli

The goal of this study is to quantify the ability to detect a change in formant frequency based on changes in the responses of neurons that are tuned to the frequency of amplitude modulations. Predictions for just-noticeable-differences (jnd's) in formant frequency can be directly

compared to experimental results for human listeners [16]. In order to make this comparison, the stimuli used in the results presented here were matched to one of the sets of stimuli used in the comprehensive study of Lyzenga and Horst [16]. The results here are based on responses to a voiced sound (F0 = 100 Hz) with a single formant at 2000 Hz, created using a triangular spectral envelope with slopes of 200 Hz/octave (Fig. 3). Lyzenga and Horst results showed that listeners had patterns of discrimination thresholds for stimuli with simple triangular spectral envelopes that were similar to those for more complex spectral envelopes that were designed to match the detailed spectral envelope of formants in actual speech sounds.

Fig. 3. Single-formant vowel-like sounds are shown with two types of spectral envelopes that were studied by Lyzenga and Horst [16]. In both cases,



the underlying structure of the sound is a set of harmonics of the fundamental frequency, or pitch. The amplitudes of the harmonics were either gradually varied across frequency, using amplitudes computed by a Klatt synthesizer, or they were varied according to a simple triangular spectral envelope. Stimuli created with a triangular spectral envelope were used for the results presented here. Lyzenga and Horst [16] described differences in jnd for triangular (or more complex) stimuli that had the spectral peak aligned with one harmonic (top) or positioned between two harmonics (bottom). Listeners were less sensitive when the spectral peak was aligned with a harmonic frequency (see text). (Adapted from [16] Fig. 1, k and l).

B. Neural Models

Two neural models were used for the calculations presented here (Fig. 4). A computational model for the auditory periphery [4] was used to simulate a population of AN responses. This model includes the sound-level-dependent bandwidth and gain of frequency-tuned cochlear responses, rate adaptation, rate saturation, and frequency-dependent phase-locking. In particular, this AN model makes accurate predictions of the responses of AN fibers to signals with fluctuating amplitudes [4].

Responses of midbrain neurons in the IC were simulated using the same-frequency inhibitory-excitatory (SFIE) model of Nelson and Carney [8] (Fig. 5). This model explains the tuning of IC neurons to the frequency of amplitude modulations. AM tuning is achieved by the interplay between relatively sluggish inhibitory responses and relatively fast excitatory responses.

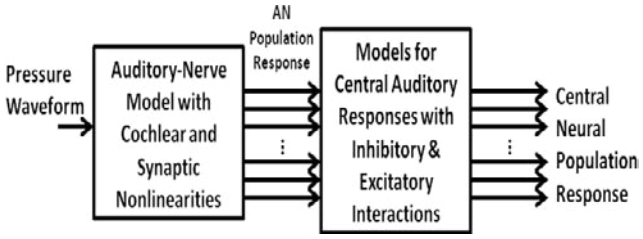


Fig. 4. Schematic of models used for the predictions presented in this study.

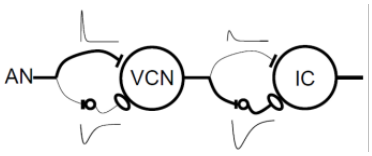


Fig. 5. Schematic illustration of the Same-Frequency Inhibitory-Excitatory (SFIE) model for AM tuning of IC neurons [8]. One of these models was used for each audio frequency channel in the simulations presented here. All SFIE models had a BMF of 100 Hz, which was equal to the fundamental frequency of the vowel-like sound that was used as the input waveform. (Simulation of the entire IC population would require sets of SFIE models with the entire range of BMFs for each audio-frequency channel.)

C. Calculating the just-noticeable difference (jnd)

The strategy for computing the jnd for formant frequency discrimination was based on the approach of Siebert [17]-[19] and Heinz *et al*. [20] for one-parameter discrimination of auditory stimuli. The developed calculations based on the Cramer-Rao bound, or equivalently a likelihood ratio test (see [20]), using the responses of a model for a population of AN fibers. In the case of the study presented here, the one parameter being manipulated was the peak of the triangular spectral envelope. In addition, rather than making predictions based on changes in the rate or timing of model AN responses, the predictions presented here are based on the responses of a population of model midbrain neurons that have band-pass tuning for amplitude-modulation frequency. It should be noted that although a single parameter is manipulated when the frequency of the spectral peak is changed, the amplitude of all of the harmonics change as a result. Nevertheless, the following calculation combines the information present across time and across the population of fibers to derive a single value for the predicted jnd for formant frequency.

The jnd is inversely proportional to the information in the responses of each neuron in the population, and this information is related to the change in rate (or in the timing pattern of the response) normalized by the variance. For the common assumption of Poisson variance in the neural responses, the variance is approximated by the mean rate. Thus, jnd is calculated as

$$\Delta F_{JND} = \left(\sum_i \int_0^T \frac{1}{r_i(t,F)} \left[\frac{\partial r_i(t,F)}{\partial F} \right]^2 dt \right)^{-1/2} \qquad (1)$$

where F is the peak frequency of the spectral envelope (see Fig. 3), and r_i is the time-varying discharge rate of the i^{th} neuron in the population (see Eq 3.1 in [20]). The calculation can be made using the entire time-varying rate function, as shown in the equation above; this result is referred to as the all-information prediction because it is based on both rate and timing information in the neural responses. Alternatively, predictions can be made based only on average rate information, by first averaging the discharge rate over the stimulus duration, thus discarding detailed temporal information [20]. Computationally, the jnd's can be computed by finding model responses to slightly different stimuli (see Fig. 6 below). Then the point-by-point difference between the two responses, normalized by the size of the increment in the parameter that varied between the two stimuli, provides an approximation to the partial derivative in Eq. 1. The results are then combined over time (for the all-information estimate) and over the population of model neurons, as in Eq. 1. For further details about the computation of jnd from model population responses, see [1] and [20]. Code for both the AN and SFIE models is available at: http://www.urmc.rochester.edu/labs/Carney-Lab/publications/auditory-models.cfm .

III. Results

Fig. 6 illustrates population responses for model IC (top) and AN (bottom) neurons. Each plot shows population responses to two stimuli that differed in peak frequency by 1 Hz; the harmonic frequencies do not differ across the two stimuli, but the amplitude of each harmonic in the stimulus was affected by the slight difference in the frequency of the spectral peak. Populations consisted of 100 model neurons tuned to best frequencies that were logarithmically spaced over 2 octaves surrounding the stimulus peak frequency. Fig. 6A shows the population response to a stimulus in which the spectral peak (2000 Hz) was aligned with a harmonic frequency (F0 = 100 Hz); Fig. 6B is for a stimulus in which the peak frequency (2050 Hz) fell between two harmonics. The population responses show discharge rates averaged over the time course of 500-ms duration stimuli with the triangular spectra shown in Fig. 3. The maximum amplitude of the spectral peak was 60 dB SPL, and the overall rms values of the two stimuli were matched. The jnd's were calculated based on differences in the model responses for neurons tuned near the spectral peak (dark lines in Fig. 6, also these population subsets are enlarged in the insets).

There are interesting qualitative differences between the two population responses due to the difference in alignment of the peak of the spectral envelope and the harmonic frequencies. In Fig. 6A, the AN responses near the harmonic frequency that is aligned with the peak are the least modulated (see Fig. 2A), and thus the model IC cells tuned near the spectral peak have strongly *reduced* responses. In Fig. 6B, there is no single dominant harmonic; as a result, the AN responses have larger amplitude modulations in general, resulting in higher rates in responses of the IC neurons. In addition, there are *two* notches in the IC population response; these notches are at the locations of the two harmonics that straddle the peak of the spectral envelope. Note that in both cases the AN population responses are characterized by a single broad peak.

The jnd calculated for the IC population responses in Fig. 6A was 9.4 Hz and for Fig. 6B it was 7.5 Hz. Smaller jnd's for stimuli in which the spectral peak fell between two harmonics were also observed for human listeners. For human subjects with normal hearing, the jnd was 0.6%, or 12 Hz for a peak at 2000 Hz, when the spectral peak was aligned with a harmonic [16]. In comparison, the jnd was 0.2%, or 4 Hz for a 2050 Hz peak, when the spectral peak was positioned between two harmonics [16]. Thus, the model jnd's have comparable sizes and follow a similar trend as the human data, although the difference between the two conditions was larger for human listeners than for the model calculations. The presence of two notches in the population response for the mis-aligned spectral peak (Fig. 6B) provides more features for discrimination, contributing to the lower jnd; however, the larger rates associated with the more strongly modulated

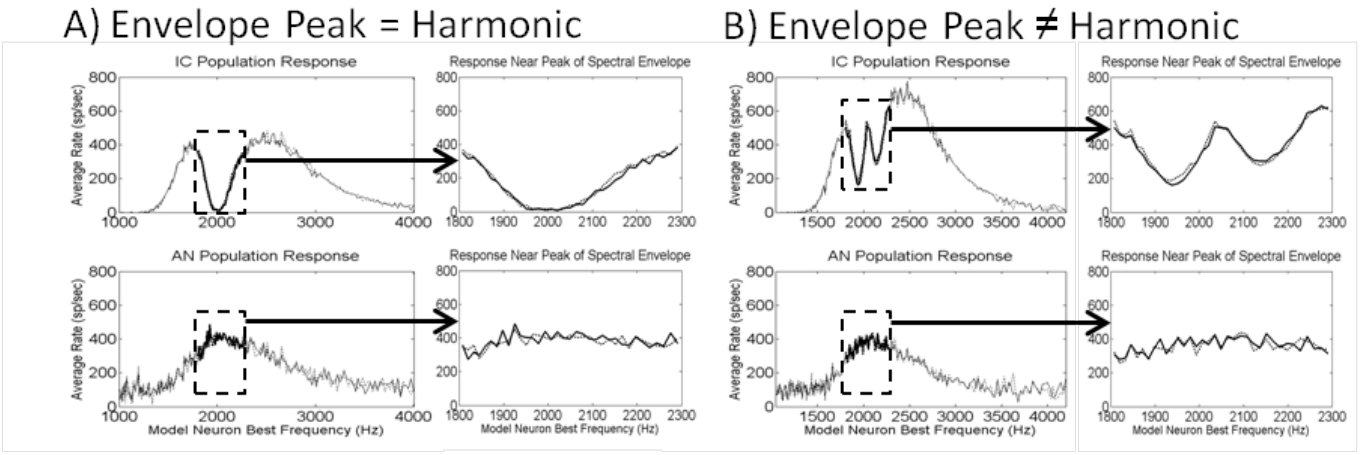


Fig. 6 – A) Population responses for model IC neurons (top) and AN fibers (bottom) for responses to two stimuli with triangular spectral envelopes, one with peak frequency = 2000 Hz (solid) and one with peak frequency = 2001 Hz (dashed). B) Responses for stimuli with peak spectral envelopes that are positioned at 2050 Hz (solid) and 2051 Hz (dashed), which fall between stimulus harmonics.

stimulus mitigate this effect somewhat, because larger rates are associated with higher variance, given the Poisson assumption (see Eq. 1). The precise values of the model jnd's depend upon detailed choices of the parameters used to set up the population responses and are being further explored in ongoing work. In addition, secondary features derived from the population responses, such as local response gradients, should be evaluated in the context of the formant-frequency discrimination problem.

IV. Conclusion

A long-term goal of this work is to understand how neural constraints on formant-frequency discrimination influence the representation of speech sounds. Languages may vary in the number of vowel phonemes or contrasts (e.g. Spanish has 5 vowels, and English has 13). However, cross-linguistically, vowels systematically disperse themselves within an acoustic/auditory vowel space defined primarily by the first and second formants or spectral energy bands (Fig. 7), which are orthogonal to the fundamental frequency, F0. The position of these two formants identifies the vowel; for some vowels (e.g. */a/* and */o/*), these formants lie on top of each other. Vowel dispersion within the space defined by the two formant frequencies has been modeled using distance metrics adjusted to reflect the actual distributions found in vowel systems (reviewed in [21]).

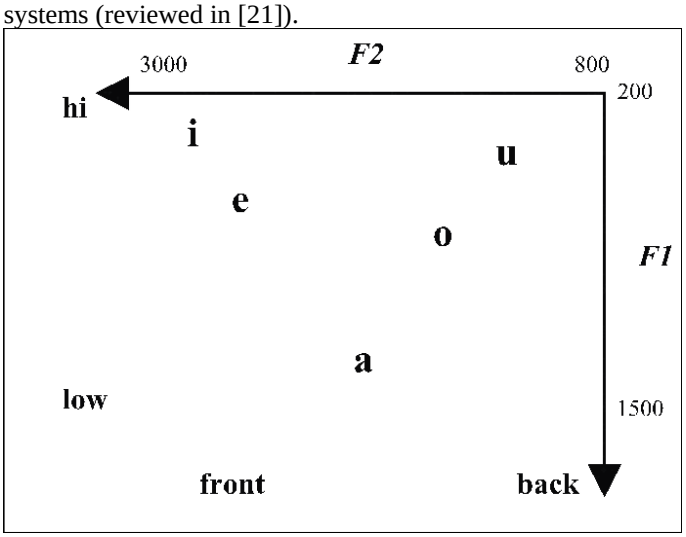


Fig. 7 - A canonical 5 vowel system exemplifying vowel dispersal in the vowel space, determined by the frequencies of the lowest two formants, F1 and F2.

The resolution for discriminations made within the vowel space is constrained by the resolution for discriminating single formants. The results presented here represent an effort to quantify the resolution within the vowel space based on the response properties of auditory neurons.

References

[1] Q. Tan and L. H. Carney, "Encoding of vowel-like sounds in the auditory-nerve: Model predictions of discrimination performance," *J. Acoust. Soc. Am.*, vol. 117, 2005, pp 1210-1222.

[2] Q. Tan and L. H. Carney, "Predictions of Formant-Frequency Discrimination in Noise Based on Model Auditory-Nerve Responses. *J. Acoust. Soc. Am.* vol. 120, 2006, pp,1435-1445.

[3] A. Palmer and S. Shamma, "Physiological representations of speech," In *Speech Processing in the Auditory System*, S.Greenberg, W. A. Ainsworth, A. N. Popper and R. R. Fay, Eds., Springer, 2004, pp. 163-230.

[4] M. S. A. Zilany, I. C. Bruce, P. C. Nelson and L.H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics," *J. Acoust. Soc. Am.*, vol. 126, 2009, pp. 2390-2412.

[5] G. Langner and C. E. Schreiner (1988) "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms," *J. Neurophysiol.*, vol. 60, 1988, pp. 1799-1822.

[6] B. S. Krishna and M. N. Semple, "Auditory Temporal Processing: Responses to Sinusoidally Amplitude-Modulated Tones in the Inferior Colliculus," *J. Neurophysiol.*, vol. 84, 2000, pp. 255–273.

[7] P. C. Nelson and L. H. Carney, "Rate and timing cues for neural detection and discrimination of amplitude-modulated tones in the awake rabbit inferior colliculus," *J. Neurophysiol.*, vol. 97, 2007, pp. 522-539.

[8] P. C. Nelson and L. H. Carney, "A phenomenological model of peripheral and central neural responses to amplitude-modulated tones," *J. Acoust. Soc. Am.*, vol. 116, 2004, pp. 2173-2186.

[9] G. Langner, "Neuronal mechanisms for pitch analysis in the time domain," *Exp. Brain. Res.*, vol. 44, 1981, pp. 450-454.

[10] M. J. Hewitt and R. Meddis, "A computer model of amplitude-modulation sensitivity of single units in the inferior colliculus," *J. Acoust. Soc. Am.*, vol. 95, 1994, pp. 2145-2159.

[11] K. Voutsas, G. Langner, J. Adamy and M. Ochse (2005) "A brain-like neural network for periodicity analysis," I*EEE Tran Sys, Man, and Cyber – Part B Cyber*, vol 35, 2005, pp. 12-22.

[12] U. Dicke, S. E. Ewert, T. Dau and B. Kollmeier, "A neural circuit transforming temporal periodicity information into a rate-based representation in the mammalian auditory system," *J. Acoust. Soc. Am.*, vol. 121, 1997, pp. 310-326.

[13] K. A. Davis, K. E. Hancock and B. Delgutte, "Computational Models of Inferior Colliculus Neurons," In *Computational Models of the Auditory System*, R. Meddis, E. Lopez-Poveda, R.R. Fay and A. N. Popper, Eds. Springer: New York, 2010, pp. 129-176.

[14] S. Baumann, T. D. Griffiths, L. Sun, C. I. Petkov, A. Thiele and A. Rees, "Orthogonal representation of sound dimensions in the primate midbrain." *Nat Neurosci*., vol. 14, 2011, pp. 423-425.

[15] L. Las, E. A. Stern and I. Nelken, "Representation of tone in fluctuating maskers in the ascending auditory system." *J. Neurosci.*, vol. 25, 2005, pp. 1503-1513.

[16] J. Lyzenga and J. W. Horst, "Frequency discrimination of stylized synthetic vowels with a single formant," J. Acoust. Soc. Am., vol. 102, 1997, pp. 755-1767.

[17] W. M. Siebert, "Some implication of the stochastic behavior of primary auditory neurons," *Kybernetik*, vol. 2, 1965, pp. 206-215.

[18] W. M. Siebert, "Stimulus transformations in the peripheral auditory system," In: *Recognizing Patterns*, P.A. Kolers and M. Eden, Eds., MIT Press, Cambridge, MA, 1968, pp. 104-133.

[19] W. M. Siebert, "Frequency discrimination in the auditory system: place or periodicity mechanisms?," *Proc. IEEE*, vol. 58, 1970, pp. 723-730.

[20] M. G. Heinz, H. S. Colburn and L. H. Carney, "Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve," *Neural Computation*, vol. 13, 2001, pp. 2273-2316.

[21] R. Diehl and B. Lindblom "Explaining the structure of feature and phoneme inventories: The role of auditory distinctiveness" in *Speech Processing in the Auditory System*. S. Greenberg, W. A. Ainsworth, A. N. Popper and R. R. Fay, Eds., 2004, Springer, pp. 101-162.