# Recurrent timing nets for auditory scene analysis

Peter Cariani
Eaton Peabody Laboratory of Auditory Physiology
Massachusetts Eye & Ear Infirmary
243 Charles St., Boston, MA 02114 USA

*Abstract.* **We have recently proposed neural timing networks that operate on temporal fine structure of inputs to build up and separate periodic signals with different fundamental periods (Neural Networks, 14: 737-753, 2001). Simple recurrent nets consist of arrays of coincidence detectors fed by common input lines and conduction delay loops of different recurrence times. Short-term facilitation amplifies correlations between input and loop signals to amplify periodic patterns and segregate those with different periods, thereby allowing constituent waveforms to be recovered. Timing nets constitute a new, general strategy for scene analysis that builds up correlational invariances rather than feature-based labeling, segregation and binding of channels.**

## I. PITCH AND AUDITORY SCENE ANALYSIS

Perhaps the most basic function of a perceptual system is to coherently organize the incoming flux of sensory information into separate stable objects [1-4]. In hearing, sound components are fused into unified objects, streams and voices that exhibit perceptual attributes, such as pitch, timbre, loudness, and location. Common periodicity, temporal proximity (onset, duration, offset), frequency, amplitude dynamics, phase coherence, and location in auditory space are some of the factors that contribute to fusions and separations of sounds.

For concurrent sounds, common harmonic structure plays perhaps the strongest role in forming unified objects and separating them [4, 5]. Harmonic complexes with different fundamentals produce strong pitches at their fundamentals, which can be heard even when the fundamental is "missing" from the power spectrum. As a rule of thumb, voices and musical instruments having the same fundamentals fuse, while those with fundamentals differing by more than a semitone (6%) can usually be separated. The mechanisms underlying pitch perception and auditory object formation therefore appear to intimately linked. One possibility is that the auditory system labels and segregates frequency channels according to pitch-related features (e.g. [6, 39]) and then binds them together at some later stage. Another possibility is that auditory objects are formed from the fine time structure of the acoustic stimulus and phase-locked neural responses. Objects would be formed from temporal pattern invariances (phase coherences) whose period would then determine the perceived pitch of the object. In this view, object formation precedes analyses of object properties (pitch, timbre, loudness, location) (see [7, 8]). Recurrent neural timing nets are demonstrations of how the latter strategy for scene analysis might be implemented neurally.

We cannot presently determine which strategy is used by real auditory systems because the central auditory mechanisms by which different voices with different fundamentals can be heard out are not yet well understood. In part this is due to the absence of a compelling theory of how pitch is represented and processed above the level of the auditory midbrain. Such a theory would need to account for pitches produced by both pure and complex tones and would need to explain how very fine pitch distinctions (< 1% in frequency) over very large dynamic ranges (> 80 dB) can be realized using neural elements whose responses are relatively coarsely tuned and highly level-dependent. Below the level of the midbrain, a large body of neurophysiological evidence [9, 10] and neurocomputational demonstration [11-16] does strongly suggest that the auditory system uses interspike interval information for pitch perception. For periodicities below roughly 5 kHz, the acoustic stimulus impresses its temporal structure on the temporal discharge patterns of auditory nerve fibers. The timings of individual spikes faithfully represent the phase structure of the stimulus, and the interspike intervals formed as a consequence of such phase-locked spike timings form an autocorrelation-like representation of the stimulus. Given that the pitches of low frequency pure tones and complex tones appear to be based on fine timing information, it is therefore entirely conceivable that stable auditory images are formed from the fine temporal structure of neural discharges (e.g. [17]). Separation of different periodic temporal patterns would be carried out on the basis of the coherence of temporal patterns by amplifying those patterns that recur in the stimulus (and consequently in neural activity patterns). A subsequent autocorrelation-like analysis based on interspike interval information would be carried out to subserve perception of the pitches and timbres of separated auditory objects.

A central unsolved problem in auditory neurophysiology concerns how the central auditory system makes use of this superabundant peripheral fine timing information for sound separation and analysis. The fields of computational neuroscience, neural networks, and signal processing can make useful contributions to the

This work was supported by NSF-EIA-BITS-013807

solution of this problem by formulating neural architectures and computational operations that can use fine timing (phase) information to do auditory scene analysis and pitch perception. Neurocomputational models then provide guides for finding neural populations in the auditory pathway that have the response characteristics needed to realize these functionalities.

The current prevailing view among auditory neurophysiologists is that a time-to-place transformation is effected in the auditory pathway somewhere between the auditory nerve and the auditory cortex. In neural network terms, time-delay neural networks are thought to convert temporal input patterns to spatialized output patterns that are then analyzed more centrally by connectionist (Hopfield) networks that operate on average firing rates. J. C. R. Licklider's original temporal autocorrelation network was an early time-delay neural network (TDNN) architecture [11, 12] that converted peripheral timing information to a spatial pattern of activity that represented the autocorrelation function of the stimulus (so as to account for the autocorrelation-like character of pitch perception). More recent proposals have involved tuned periodicity detectors [18], but unfortunately these modulation-tuned elements do not carry out the right kinds of operations for either pitch analysis [10] or object separation. Thus far, no neural temporal autocorrelators, true pitch detectors, spectral pattern analyzers, or, for that matter, *any* strong neural correlates of the pitches of complex tones have been found above the midbrain.

For these reasons, we have strived to develop new heuristics for how auditory images might be formed and separated. One route is to explore alternative kinds of signal transformations[19] that would utilize temporal pattern information in novel, unforeseen ways. Another possible general solution to the problem is to keep the information in the time domain and to use mechanisms for temporal processing to form and then analyze auditory objects. This is the strategy outlined here. As an alternative to time-delay transformations, simple networks that operate on temporally-structured inputs to produce temporally-structured outputs were conceived and called "neural timing nets." Both feedforward and recurrent networks were considered and their basic computational properties were explored [20, 21].

## II. RECURRENT TIMING NETS

Recurrent timing networks were inspired in different ways by models of stabilized auditory images [17], neural loop models [22], adaptive timing nets [23], adaptive resonance circuits [24], the precision of echoic memory, and the psychology of temporal expectation [25, 26]. Although much is known about time courses of temporal integration that are related to auditory percepts (pitch, timbre, loudness, location, object separation, and various masking effects), few neurocomputational models exist for how incoming information in the auditory periphery is integrated over time by the central auditory system to form stabilized auditory percepts. If the incoming information is indeed encoded in temporal patterns of spikes, then it is not unreasonable to consider possible neural architectures that store temporal patterns in (centrally disinhibited) reverberating circuits. One envisions the signals themselves circulating in closed transmission loops or regenerated via cellular recovery mechanisms. Reverberating temporal memory traces would be compared with incoming patterns via coincidence-detectors that compute temporal correlations. Neural representations would thus build up over time, dynamically creating sets of perceptual expectations that could either be confirmed or violated. Periodic signals, such as isochronous rhythms, would create the strongest temporal expectancies [27, 28].

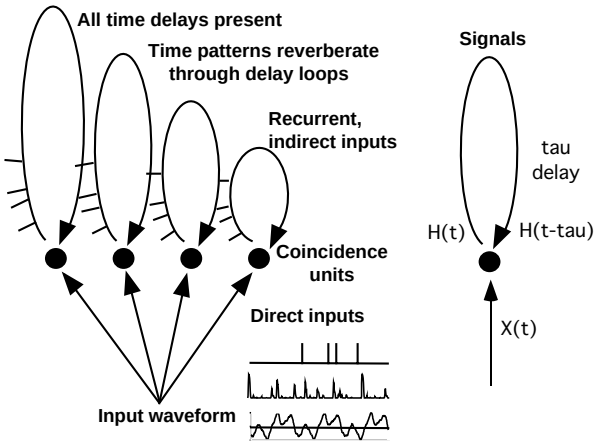


Figure 1. Simple recurrent timing net.

The simplest recurrent timing networks imaginable in these terms consist of a 1-D array of coincidence detectors having common direct inputs (Figure 1). The output of each coincidence element is fed into a recurrent delay line such that the output of the element at time t circulates through the line and arrives tau milliseconds later (the signal that arrives back at time t is the one that was emitted at t – tau). A processing rule governs the interaction of direct and circulating inputs.

In their development the networks have evolved from simple to more complex. In the first simulations [20], binary pulse trains (resembling spike trains) with repeated, randomly selected pulse patterns (e.g. 100101011-100101011-100101011...) were passed through the network. For each time step, incoming binary pulses were multiplied by variable-amplitude pulses arriving through the delay loop. In the absence of a coincidence with a circulating pulse, the input pulse was fed into the delay loop without facilitation. When coincidences between incoming and circulating pulses occurred, the amplitude of the circulating pulse was increased by 5% and the pulse was fed back into the loop. It was quickly realized that

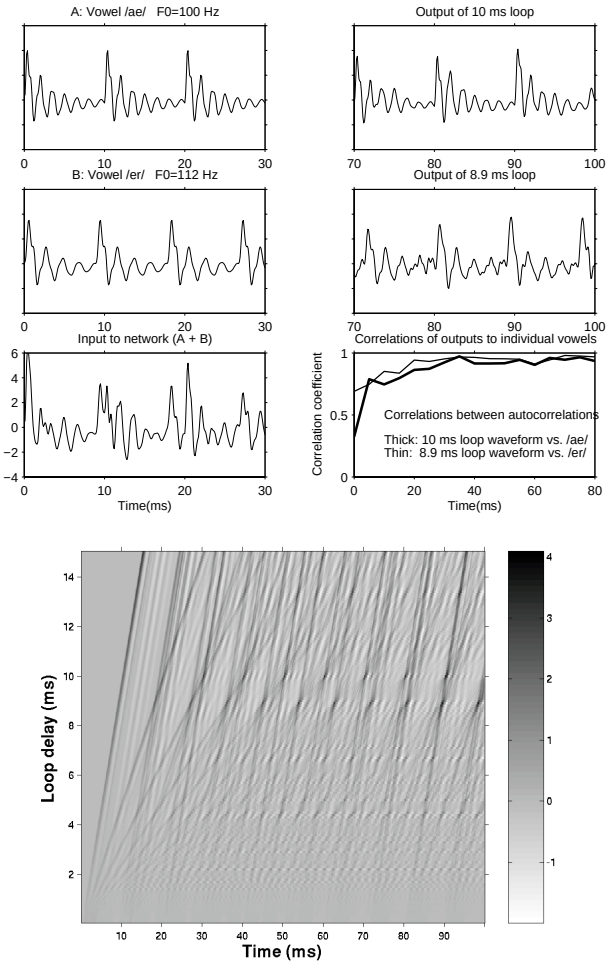
such networks rapidly build up any periodic pulse patterns in their inputs, even if these patterns are embedded amidst many other pulses. A periodic pattern invariably builds up in the delay loop whose recurrence time matches its repetition time. Thus, recurrent time patterns are repeatedly correlated with themselves to build up to detection thresholds. In effect, these autocorrelating loops dynamically create matched filters from repeating temporal patterns in the stimulus. In this manner, temporal-pattern invariances are enhanced relative to uncorrelated patterns – the network functions as a pattern-amplifier. When two repeating temporal patterns each with its own repetition period were summed together and presented to such nets, the two patterns emerged in the two different delay loops that had recurrence times that corresponded to the repetition periods of the patterns. Although the proportional facilitation rule distorted signal amplitudes, the temporal patterns of pulses corresponding to the two rhythms could be recovered in the circulating waveforms. A neural network can therefore carry out an analog-style separation of signals in the time-domain. To do this, inputs need to be temporally coded, processing elements must have sufficiently narrow coincidence windows, delays must be relatively precise, and processing rules must be judiciously chosen.

## III. SEPARATION OF DOUBLE VOWELS

Although binary pulse trains resemble spike trains of individual neurons, most real neural information processing appears to be carried out by large ensembles of neurons working in concert. Subsequent simulations [21] therefore used positive real-valued input signals that qualitatively resemble neural post-stimulus time histograms (e.g. time series of spike counts that would be produced by an ensemble of similar neural elements whose discharges were stimulus-locked). Proportional facilitation was replaced by a processing rule that adaptively adjusts the output signal in a more graceful and less distorting manner.

Double vowel stimuli were used for processing because a considerable body of psychophysical, neurophysiological, and neurocomputational work had been carried out on their perception, e.g. [6, 29-35]. Double vowels are concurrently-presented pairs of vowels (Fig. 2, left panels). Human listeners are able to use relatively small differences in fundamental frequency to better separate and identify the two individual vowels. Perceptually, differences in timbre distinguish different vowel classes (e.g. /ae/ vs /er/). Most models to date (e.g. [6, 39]) have attempted with varying degrees of success and generality to segregate the two vowels by segmenting subsets of frequency channels using response features related to fundamental frequency (F0).

Pairs of synthetic vowels (double-vowels) were half-wave rectified and presented to a 1-D network (the positive




Figs. 2 & 3. Separation of double vowels.

parts of the waveforms of Fig. 2). As with the pulse trains, the constituent signals were separated into their respective delay loops. The patterns of signals in the loops resembled those shown in Fig. 3 – the loops with the strongest signals had recurrence times that corresponded to the fundamental periods of the vowels. Examination of the waveforms in these most-activated loops showed waveforms that resembled the temporal structure of the individual vowels, such that they could be recovered.

Double vowels were also processed through an auditory nerve front-end (24 frequency channels, coarse band-pass tuning, rectification, 5 kHz roll-off of phase-locking, rate-level compression, 10 kHz sampling rate) to an array of delay loops (150 per frequency). Autocorrelations of circulating waveforms in corresponding delay channels were combined across frequencies (analogous to population-interval distributions that form neural representations of pitch and timbre [9, 10]).

In both single- and multi-channel cases, when vowel fundamentals were separated by a semitone or more, the autocorrelations (and hence, power spectra) of the constituent vowels could be accurately recovered. Quality

of the separations improved as a function of ΔF0 and vowel duration. This simulation demonstrated how recurrent timing nets could be scaled up to process multichannel positive, real-valued signals similar to ensembles of auditory nerve fiber spike trains. It also showed that auditory objects can be separated even when they activate the same sets of broadly tuned frequency channels. Information related to multiple auditory objects (the two vowels) is embedded/multiplexed in the phase structure of the stimulus and phase-locked spike timings. The networks demonstrate how an auditory scene analysis system can exploit phase-coherence and F0-differences without first carrying out explicit estimations of F0 and segregating frequency channels on that basis. For example, Wang & Brown [39] employed a feature- and channel-based sorting strategy in which fine timing was used to compute channel autocorrelations for each of many narrowly tuned frequency channels. F0-related autocorrelation features were subsequently used by a synchronizing oscillator array to group the frequency channels. Although neural timing nets implicitly have processing lags due to loop delays, no explicit autocorrelation profiles are computed and analyzed, and the separated signals remain as time-series waveforms.

## IV. CURRENT IMPLEMENTATIONS

It would be useful to separate signals that are not half-wave rectified. In the interest of development of more practical signal-processing applications, we adopted a simple error-adjustment processing rule (1,2) that can operate on signals with both positive and negative values.

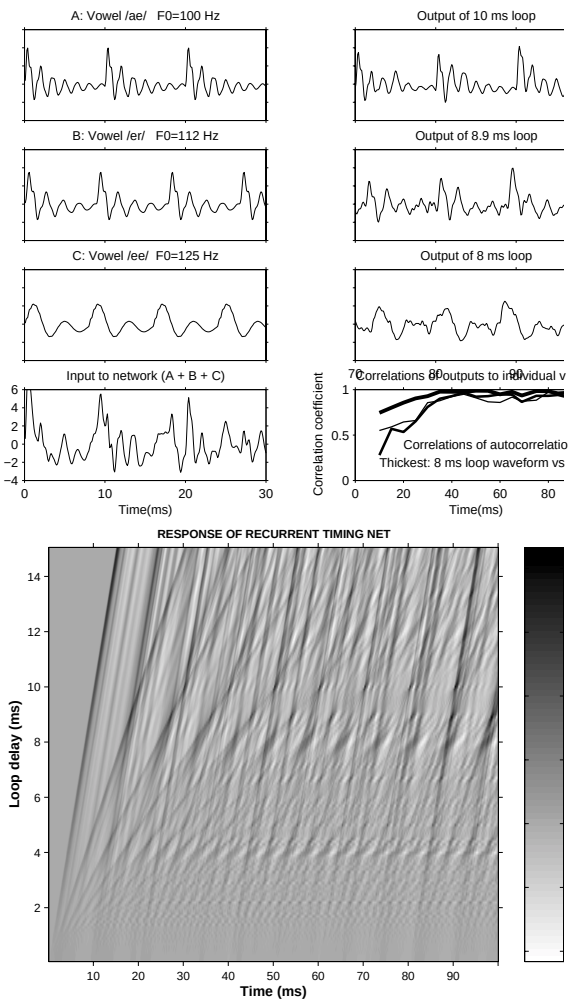
$$H(t) = H(t\text{-}tau) + B_{tau}[X(t)\text{-}H(t\text{-}tau)] \quad (1)$$

$$B_{tau} = tau/33 \text{ ms} \quad (2)$$

For each loop with recurrence time tau, at time t, X(t) is the direct input signal, H(t-tau) is the incoming circulating signal, and H(t) is the outgoing circulating signal (Fig. 1). $B_{tau}$ determines the rate of adjustment, and its dependence on tau ensures that shorter loops are not favored.

Synthetic, three-formant double vowels (/ae/, /er/) with different fundamentals (100, 112 Hz) were summed and processed by the network (Figs. 2 & 3). The signals circulating in the 150 delay loops are shown in the response map of Fig. 3, where it can be seen that the recurrence times of the loops with the highest average signal strength correspond to the periods of the two vowels (8.9 & 10 ms). The signals circulating in these two delay channels after 70 ms of processing highly resemble the two vowel constituents (Fig. 2, top four panels). Correlations between the autocorrelations of these processed signals and those of the individual vowels show how the signal separation unfolds over processing time.

One can ask how well these networks handle more than two auditory objects. A third vowel /ee/ with yet a different fundamental (125 Hz) was added to the mixture.




Figs. 4 & 5. Separation of triple vowels.

This is akin to hearing out three different kinds of musical instruments playing different notes (first that there are three different notes, second that instruments with different timbres are playing the three notes). Processing by the network resulted in the appearance of another strong signal in the response map (Fig. 5, 8 ms delay loop). Separation of the signals in the three-vowel case was somewhat slower than for two vowels (Fig. 4, bottom right), but there was only a slight reduction in the final quality of the separated signals. Network performance therefore appears robust. One caveat is that we have used synthetic vowels with an unvarying fundamental period. Processing of natural signals would require some (pitch) tracking across delay loops, which would likely reduce correlations. Nevertheless, the correlation-based nature of the processing makes for a very transparent representation that can accommodate multiple auditory objects with overlapping power spectra. Use of fine time structure permits information about multiple objects to be multiplexed in the same (neural, frequency) channels in a manner that minimizes destructive interference.

## V. ENHANCEMENT OF VOWELS IN NOISE

A possible use of recurrent timing nets is to enhance periodic sounds in noisy environments. Such processing would be useful for processing music and voiced speech. Reductions in effective S/N ratios could be expected to improve speech reception by human listeners and automatic recognition by machines. Related kinds of correlation-based strategies were used in the 1950's to detect periodic signals in noise [36, 37], in situations where the period of the target signal was known *a priori*. The present networks systematically sample all possible delays, such that the optimum delay(s) can be determined by choosing the loop(s) with the largest signal rms.

In order to assess the performance of the network in noise, a synthetic, three-formant vowel (/ae/, F0=100 Hz) was added to frozen white noise at different S/N ratios (S/N = -20-20 dB). The input and output signals from the optimum delay loop (tau = 10 ms) are shown in Figure 6 (top panels). Correlations between the autocorrelation of these signals and that of the vowel in near-quiet (20 dB S/N) are shown in bottom panels. Similarities between the processed signals and the minimal-noise case improve with S/N and processing time. For all S/N ratios below 1, the network produced output signals (thick curves) that had higher correlations than the unprocessed, input signals (thin curves). Processing by the network shifts the curves to the left, an improvement in S/N of roughly 4-10 dB.

## VI. BROADER IMPLICATIONS

Recurrent timing nets use the periodic patterns in their inputs to dynamically form matched templates that they compare with subsequent inputs. For ease of visualizing their behavior, we have considered ordered arrays of monosynaptic delay loops. It is conceivable that such processing can also be carried out in randomly connected networks, provided that recurrent, multisynaptic pathways are available that span a wide range of loop-delays. If coincidence elements that are transiently facilitated by temporal coincidences are used, then it is not hard to envision how timing nets might support dynamically-formed reverberatory memories capable of retaining temporal patterns and interspike interval statistics. Such networks would be akin to self-organizing recurrent synfire chains [38] in which both synchrony and temporal patterning of spikes play critical roles.

Obviously whether the abstract nets discussed here resemble those operant in real brains will only become clear when the specific neural representations and processing mechanisms are much more firmly understood. Nevertheless, knowing what kinds of neural mechanisms to look for is essential for reverse-engineering how brains work. Timing networks add to that growing list of candidate mechanisms.

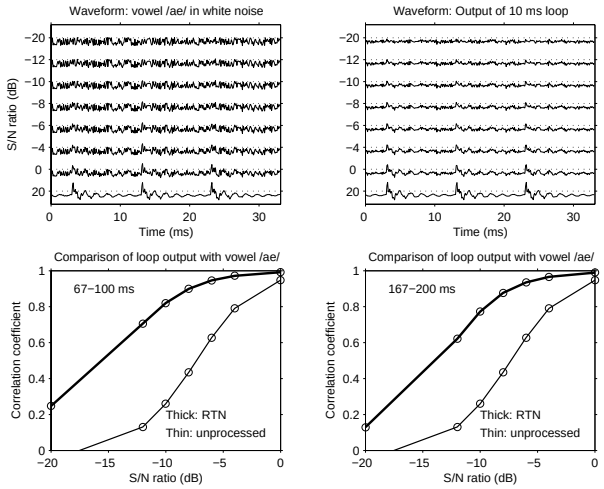



Fig. 6. Performance of the network in noise.


## ACKNOWLEDGMENTS

We thank Ramdas Kumaresan, Alan Lingren, Jesse Hanson, and Yadong Wang of the University of Rhode Island for their signal processing insights.


## REFERENCES


[1] A. S. Bregman, "Asking the "what for" question in auditory perception," in *Perceptual Organization*, M. Kubovy and J. R. Pomerantz, Eds. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1981, pp. 99-118.

[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[3] S. Handel, *Listening*. Cambridge: MIT Press, 1989.

[4] D. K. Mellinger and B. M. Mont-Reynaud, "Scene analysis," in *Auditory Computation*, H. Hawkins, T. McMullin, A. N. Popper, and R. R. Fay, Eds. New York: Springer Verlag, 1996, pp. 271-331.

[5] W. M. Hartmann, "Pitch perception and the segregation and Integration of auditory entities," in *Auditory Function: Neurobiological Bases of Hearing*, G. M. Edelman, Ed. New York: John Wiley & Sons, 1988, pp. 623-347.

[6] R. Meddis and M. J. Hewitt, "Modeling the perception of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 91, pp. 233-245, 1992.

[7] M. Kubovy, "Concurrent-pitch segregation and the theory of indispensible attributes," in *Perceptual Organization*, M. Kubovy and J. R. Pomerantz, Eds. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1981, pp. 55-98.

[8] C. J. Darwin and R. B. Gardner, "Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality," *J Acoust Soc Am*, vol. 79, pp. 838-45, 1986.

[9] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. I. Pitch and pitch

salience. II. Pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch.," *J. Neurophysiology*, vol. 76, pp. 1698-1734, 1996.

[10] P. Cariani, "Temporal coding of periodicity pitch in the auditory system: an overview," *Neural Plasticity*, vol. 6, pp. 147-172, 1999.

[11] J. C. R. Licklider, "Three auditory theories," in *Psychology: A Study of a Science. Study I. Conceptual and Systematic*, vol. Volume I. Sensory, Perceptual, and Physiological Formulations, S. Koch, Ed. New York: McGraw-Hill, 1959, pp. 41-144.

[12] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. VII, pp. 128-134, 1951.

[13] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification," *J. Acoust. Soc. Am.*, vol. 89, pp. 2866-2882, 1991.

[14] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Am.*, vol. 102, pp. 1811-1820, 1997.

[15] R. Lyon and S. Shamma, "Auditory representations of timbre and pitch," in *Auditory Computation*, H. Hawkins, T. McMullin, A. N. Popper, and R. R. Fay, Eds. New York: Springer Verlag, 1995, pp. 517.

[16] M. Slaney and R. F. Lyon, "On the importance of time - a temporal representation of sound," in *Visual Representations of Speech Signals*, M. Cooke, S. Beet, and M. Crawford, Eds. New York: John Wiley, 1993, pp. 95-118.

[17] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.*, vol. 98, pp. 1890-1894, 1995.

[18] G. Langner, "Periodicity coding in the auditory system," *Hearing Res.*, vol. 60, pp. 115-142, 1992.

[19] R. Kumaresan and Y. Wang, "On representing signals using only timing information," *J Acoust Soc Am*, vol. 110, pp. 2421-39, 2001.

[20] P. Cariani, "Neural timing nets for auditory computation," in *Computational Models of Auditory Function*, S. Greenberg and M. Slaney, Eds. Amsterdam: IOS Press, 2001, pp. 235-249.

[21] P. Cariani, "Neural timing nets," *Neural Networks*, vol. 14, pp. 737-753, 2001.

[22] R. W. Thatcher and E. R. John, *Functional Neuroscience, Vol. I. Foundations of Cognitive Processes*. Hillsdale, NJ: Lawrence Erlbaum, 1977.

[23] D. M. MacKay, "Self-organization in the time domain," in *Self-Organizing Systems 1962*, M. C. Yovitts, G. T. Jacobi, and G. D. Goldstein, Eds. Washington, D.C.: Spartan Books, 1962, pp. 37-48.

[24] S. Grossberg, *The Adaptive Brain, Vols I. and II.* New York: Elsevier, 1988.

[25] M. R. Jones, "Time, our lost dimension: toward a new theory of perception, attention, and memory," *Psychological Review*, vol. 83, pp. 323-255, 1976.

[26] R. R. Miller and R. C. Barnet, "The role of time in elementary associations," *Current Directions in Psychological Science*, vol. 2, pp. 106-111, 1993.

[27] P. Cariani, "Temporal codes, timing nets, and music perception," *J. New Music Res.*, vol. 30, pp. 107-136, 2002.

[28] P. Fraisse, "Time and rhythm perception," in *Handbook of Perception. Volume VIII. Perceptual Coding*, E. C. Carterette and M. P. Friedman, Eds. New York: Academic Press, 1978, pp. 203-254.

[29] Q. Summerfield and P. F. Assmann, "Perception of concurrent vowels: effects of harmonic misalignment and pitch-period asynchrony," *J. Acoust. Soc. Am.*, vol. 89, pp. 1364-1377, 1991.

[30] P. F. Assmann and Q. Summerfield, "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 88, pp. 680 - 697, 1990.

[31] A. de Cheveigne, "Waveform interactions and the segregation of concurrent vowels," *J Acoust Soc Am*, vol. 106, pp. 2959-72, 1999.

[32] P. Cariani and B. Delgutte, "Interspike interval distributions of auditory nerve fibers in response to concurrent vowels with same and different fundamental frequencies," *Assoc. Res. Otolaryngology. Abs.*, pp. 373, 1993.

[33] P. Cariani and B. Delgutte, "Transient changes in neural discharge patterns may enhance separation of concurrent vowels with different fundamental frequencies [Abstr]," *J. Acoust. Soc. Am.*, vol. 95, pp. 2842, 1994.

[34] A. R. Palmer, "The representation of concurrent vowels in the temporal discharge patterns of auditory nerve fibers," in *Basic Issues in Hearing*, H. Duifhuis, J. W. Horst, and H. P. Wit, Eds. London: Academic Press, 1988, pp. 244-251.

[35] A. R. Palmer, "Segregation of the responses to paired vowels in the auditory nerve of the guinea pig using autocorrelation," in *The Auditory Processing of Speech*, S. M.E.H., Ed. Berlin: Mouton de Gruyter, 1992, pp. 115 - 124.

[36] F. H. Lange, *Correlation Techniques*. Princeton: Van Nostrand, 1967.

[37] W. Meyer-Eppler, "Exhaustion methods of selecting signals from noisy backgrounds," in *Communication Theory*, W. Jackson, Ed. London: Butterworths, 1953, pp. 183-194.

[38] M. Abeles, *Corticonics.* Cambridge: Cambridge University Press, 1990.

[39] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, pp. 684-697, 1999.