Novel Speech Processing Mechanism Derived from Auditory Neocortical Circuit Analysis

Boris Aleksandrovsky, James Whitson, Gretchen Andes
Gary Lynch, and Richard Granger

Information and Computer Science Department
University of California at Irvine and
Intrinsic Circuits, Inc.
Irvine, California 92717-3800

Abstract

Analysis of the prominent anatomical and physiological features of auditory thalamus and neocortex has enabled construction of models designed to identify functionality emergent from these biological circuits. These models have recently been shown to provide powerful computational mechanisms for processing of continuous time-varying sequences such as speech; testing on speech databases has yielded positive initial results that are reported here. The model constitutes a novel hypothesis of underlying functions of auditory neocortex, and also represents a novel approach to speech processing.

1. Generalized cortical memory model

Research in our laboratory has been concentrating on the phenomenon of long-term potentiation (LTP) [3], which is the most likely candidate for a substrate of neocortical learning and memory. A set of simple learning rules was formulated based on physiological properties of LTP - i) synaptic weight can only increase, ii) every increase is small fixed change, and iii) low saturation threshold permits only 5-10 weight increases over the whole period of training [2] [10] . A series of models were constructed based on the known anatomical cortical features - sparse-random connectivity in the superficial cortical layers, emergence of the cortical patches defined by the radius of the local inhibition, and feedback inhibition and masking. The above variety of neocortical features specify a biologically constrained class of microcircuits, which typically perform pattern recognition or classification via competitive learning and lateral inhibition [6] [5] . Simulations of those circuits lead to efficient hardware implementations, with a proven utility for pattern recognition via efficient approximation of statistical pattern recognition methods (e.g. Bayes classifiers) [5].

Key anatomical properties of the auditory model being reviewed in [16] (see also Table 1) include i) topographic (MGv) versus broadly-tuned (MGm) thalamic nuclei, convergently projecting to primary auditory cortex; local cortical circuits composed of roughly 100:1 excitatory to inhibitory cells with lateral inhibition; and iii) vertical columnar organization projecting from middle to superficial to deep layers. Key physiological properties include: i) plastic (NMDA-dependent) synapses from broadly tuned MGm afferents versus non-plastic synapses from topographic MGv afferents; ii) plasticity via long-term potentiation (LTP); and iii) time courses for excitation versus inhibition of roughly 1:100. Learning in the model is based on the physiological induction and expression rules for synaptic long-term potentiation or LTP [12] [11] which have been shown in previous modeling efforts to give rise to useful computational properties [2] [10] [9] [5] [7] [8].

Superficial Cortical Layer

- Broadly tuned afferents from non-specific thalamic nucleus
- Sparse connectivity ($p \sim .1$)
- LTP of afferent synapses from non-specific nucleus
- Topographic afferents projecting vertically from the middle layer
- Topographic vertical projections from superficial to deep layer
- 100:1 ratio of excitatory to inhibitory cells

Middle Cortical Layer

- Topographic afferents from specific nucleus
- Equal number of excitatory and inhibitory cells
- Vertical projections to superficial layer
- Nonplastic synapses

Deep Cortical Layer

- Vertical afferents from superficial layer
- 100:1 ratio of excitatory to inhibitory cells
- Broadly tuned feedback to non-specific nucleus
- Topographic feedback to specific nucleus
- LTP of vertically-projecting afferents from superficial layer

Nonspecific thalamic nucleus (MGm)

- Glomerular organization
- Broadly tuned afferents from periphery
- Broadly tuned feedback from deep cortical layer
- 100:1 ratio of excitatory to inhibitory cells

Specific thalamic nucleus (MGv)

- Glomerular organization
- Topographic afferents from periphery
- Topographic feedback from deep cortical layer
- 100:1 ratio of excitatory to inhibitory cells

Receptor Time Courses

- AMPA EPSP brief (~20 msec)
- NMDA EPSP intermediate (~60 msec)
- GABA IPSP long (~100 msec)
- Feedback GABA refractory period

Table1. Biological Features of the Auditory Cortical Model

The auditory thalamic model consists of two distinct thalamic nuclei (specific and nonspecific); the cortical network model is simplified to consist of a superficial cell layer (I-III), a middle layer (IV) and a deep layer (V-VI). The specific thalamic nucleus projects topographically to the middle layer where as the nonspecific nucleus projects nontopographically (randomly) to the superficial layer. The middle layer projects vertically to the superficial layer, which in turn projects vertically down to the deep layer, forming a vertically-organized cortical "column." The middle layer, already topographically organized, is static and does not learn; the superficial and deep layers each learn via algorithms derived from the physiological rules of LTP [2] [9]. It can be seen that the superficial layer receives convergent inputs from the middle layer and from the nonspecific thalamic nucleus; learning occurs where these inputs converge. This gives rise to a form of hierarchical clustering, of a kind first identified based on modeling of the superficial layers of the olfactory cortex [2]. The deep layer learns brief feature sequences of length two, i.e., the transitions between two adjacent features.

2. Example of the model operation

This section illustrates an operation of the cortical mode [16] when presented with a stream of speech features. Multitude of speech features (see section 3 for discussion of feature extraction) are abstracted here to represent symbols "C", "A" and "T" (e.g. forming a pronunciation of word "CAT"); in the model those features represent an n-dimensional vector composed of auditory thalamic neuron firing pattern in response to speech signal.

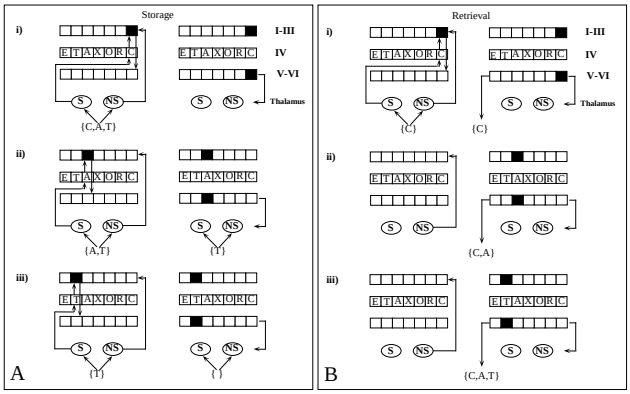

Fig1. Cortical Storage and Retrieval Operations

Following is an example of the storage operation on a stream of symbols "C,A,T" (Figure 1a):
i) The specific and nonspecific thalamic nuclei pass the first input feature ("C") to the cortex, topographically to the middle layer and nontopographically to the superficial layer, respectively. Resulting middle layer excitation pattern is propagated to the superficial layer along the limited vertical extend of the cortex. At the site(s) in the superficial layer where these inputs converge, the input is learned as a category. The category is passed to the deep layer, which learns the transition from no character to C, and outputs a random pattern in response, which is fed back to the nonspecific thalamic nucleus.
ii) The next input to cortex from the specific nucleus is the next peripheral input ("A"), whereas the input from the nonspecific nucleus is the random feedback received from the deep layer in response to the prior peripheral input (C). Thus the superficial layer now learns a category code for the random feedback from C, at the site determined by the topographic input A. The deep layer learns the transition between features C and A, and outputs a new random response which is fed back to thalamus.
iii) The process in ii) is repeated with the next peripheral input T via the specific nucleus and the random feedback from the cortex in response to A via the nonspecific nucleus. The overall storage process essentially encodes a sequence of random patterns among a collection of labeled recognition nets, with each successive pattern generated from its predecessor.

Figure 1b illustrates the retrieval of a learned cue from a unique prefix to that cue. If the sequence CAT is the only sequence beginning with the feature C that has been presented to the network, then the entire string can be retrieved simply by presentation of the unique prefix. It is worth noting that any sequence can be trained with a preappended unique code, thus enabling cued recall of the sequence via presentation of that code. Alternatively the end of the previously recognized sequence or some other contextual event can serve as a cue for successive recall of the learned sequence (see Future Work section and [1] for discussion:)
i) The unique prefix C (or a unique prefix code) is presented to thalamus and passed to cortex. The only columnar network that recognizes the input is the "C" column; the resulting recognition triggers the same random deep layer response as was previously evoked by the C input. This pattern is fed back to thalamus.
ii) No further input enters, so the feedback pattern becomes the input to cortex via the nonspecific nucleus, with no input via the specific nucleus, since no input has arrived. The signal from the nonspecific nucleus elicits a response only in the region where this signal was previously learned, i.e., the column corresponding to the input A.
iii) The deep layer response to this is then fed back to thalamus, and the next nonspecific thalamic input to cortex then elicits a response only where it was trained, which is the column corresponding to the input T. The final input from nonspecific thalamus to cortex (random code generated by T) goes unrecognized, since it was never trained, ending the retrieval of the string. Thus the sequence retrieval process amounts to the regeneration of the previously-learned chain of random patterns via the sequential activation of labeled recognition columns in cortex.

3. Application of the algorithm to speech recognition

Speech signals were preprocessed to allow for extraction of category name sequences from the speech stream in temporal order. Speech streams were split into time slices of 20 msec in duration ; the sampling rate is modeled after the cortical gamma rhythm of 40-50Hz (see Figure 2 for illustrative example).

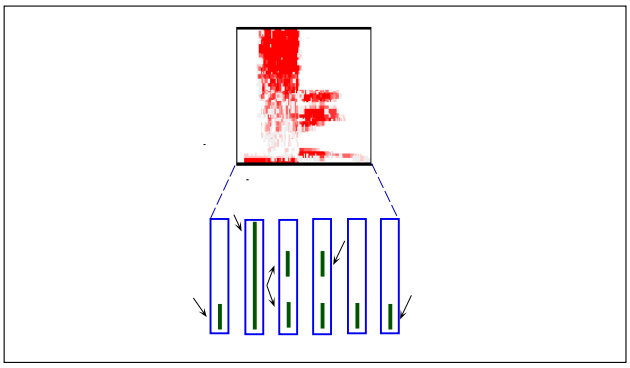

Fig2. Illustration of Speech Feature Extraction

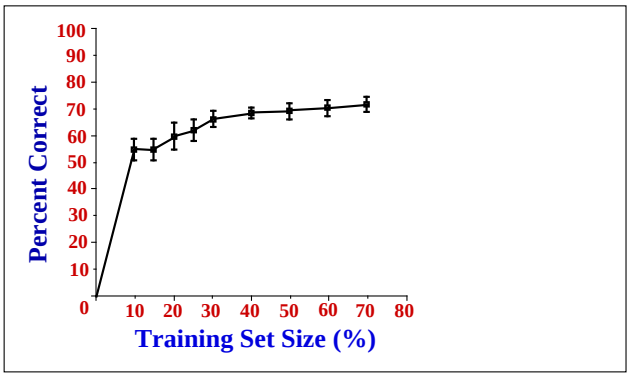

Fig3. Percent Correct Performance

In order to model peripheral auditory neuronal phenomena of adaptation, nonlinear frequency and intensity compression, equal-loudness preemphasis and onset-coding the combination of cochleogram [13] and linear predictive analysis of speech (see e.g. [14]) was constructed. Application of Lyon's cochlear model on speech streams yielded an 84-dimensional output vector per each of the time slices, each dimension corresponding to a particular frequency band and representing the rate of the hair cell firings along the extent of the cochlea. Lyon's model provides automatic gain control (intensity compression) with saturation (auditory adaptation) , represents the intensity of signal along the logarithmic frequency scale (frequency compression), and can be tuned to represent equal-loudness preemphasis (selective enhancements of middle range frequencies associated with best hearing). Output of the cochlea model was then analyzed by a linear autoregressive model modified to operate in the frequency domain; this model allows for reliable identification of transient events in speech (such as noise bursts, voice stops and onsets) via selective enhancements of those features and suppressing speech events which are extended in time (like voiced and fricative components). Resulting 6-dimensional feature vectors (one per each of the time slices) were clustered by unsupervised hierarchical clustering algorithm modelled after superficial layers of the cortex [2], which led to the abstract category name sequences, which in turn were the final peripheral input to the cortical model.

3.1. Methods and Results

The ISOLET dataset supplied by OGI speech processing group (see [4]) was used in speech processing experiments. The dataset contains 3900 utterances, each of them is the pronunciation of one of every 26 letters of the alphabet, each spoken twice by each of 150 speakers. Tests were run using cross-validation: a 10% portion of the database was used for training, and the remaining 90% for recognition testing; then a different 10% portion was used for training and the remainder for testing, etc. Figure 3 shows percent correct performance achieved by training on different data set sizes, with error bars indicating standard deviations in results on different cross-validating datasets.

Not every speaker (on average only every seventh speaker) was present in the training set, however the system had showed reliable speaker-independent recognition results of around 71%. It was worth noting that speaker-independent recognition of collections of phonemes without the context is a very difficult task even for human listeners (OGI group[15] reported top recognition rates of only 55% for both human listeners and ASR system on vowel recognition). Figure 4 presents an example of a cochleograms of letters "B" and "D". Sliding hollow bars represent transitions and numbers below them correspond to cluster names. One can see that the only difference between B2 and D2 is a single brief transition at the beginning; B1 and D1 are virtually identical ; D1 and D2 on the other hand have very little in common. All those factors invariably contributed to the difficulty of the task.

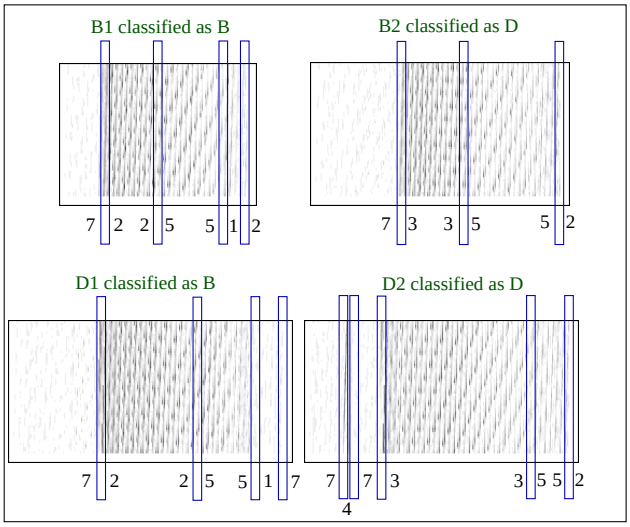

Fig4. Examples of Misclassification of Speech Data

4. Future Work

In light of application of neocortically-derived algorithms to signal processing tasks of the foremost importance are the identification of cooperative computation between different cortical areas, which in the brain facilitates context- driven perception . Initial work had been performed on modeling the mechanisms

of cortico-cortical computation, facilitated both by feedforward (from primary to secondary, to tetriary cortices) and feedback (from higher cortices to the lower) connections. Higher level cortical regions essentially learn sequences of outputs of the prior cortical regions, and feedback information about those higher-level sequences. This enables a form of "contextual" processing in which information about larger sequences (e.g., a word) is available to help disambiguate information about a shorter sequence (e.g., a syllable in that word). Initial work has shown that implementation of this hierarchical cascade of cortical areas and their corresponding cortico-cortical connectivity yields a form of *context* across multiple levels of processing (speech primitive, syllable, character, word, phrase, sentence). Further properties of time dilation, noise suppression, implicit segmentation, and associative recall arise from operation of this extended model [7], and further work in this area is ongoing.

5. Summary

Algorithm thus presented exhibits a number of useful properties in context of temporal signal processing:

- no implicit requirement for segmentation at test time, due to the ability to learn and recognize temporal sequences of fixed length.
- good speaker-independent recognition rates by extracting common speech stream features across speakers.
- capability for "time dilation", recognizing stretched and compressed versions of spoken inputs ("cat" vs. "caaat") due to implicit repeat delays in thalamocortical model.
- extendibility of the model to continious speech processing via modelling of cortico-cortical "contextual processing".

6. Conclusions

Current paper presents the novel hypothesis of the functional organization of neocortical sensory processing. Resulting architecture was shown to poses a variety of useful computational properties, and as such was successfully applied to difficult temporal signal processing tasks. Results thus obtained show promise to successfully scale to even more difficult tasks, such continuous speaker-independent speech recognition, due to no requirement for implicit segmentation and the ability to incorporate contextual information. To prove that, the same algorithm was successfully applied to OCR domain, yielding positive initial results [1]. Ongoing work includes the application of the model to continuous speech processing, as well as to the connected handwriting.

Acknowledgments

Authors are grateful to Richard Lyon of Apple Computer for making his cochlea model available and for discussions of cortical model. Authors wish to thank Mark Fanty of OGI for providing OGI tools and data.

References

[1] B. Aleksandrovsky, J. Whitson, A. Garzotto, G.Lynch, and R.Granger. An algoritm derived from thalamocortical circuitry stores and and retrieves temporal sequences. *Proceedings of ICPR*, August 1996.

[2] J. Ambros-Ingerson, R. Granger, and G. Lynch. Simulation of paleocortex performs hierarchical clustering. *Science*, 247:1344–1348, 1990.

[3] T. Bliss and T. Lomo. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology*, 232:331–356, 1973.

[4] R. Cole, M. Fanty, Y. Muthusamy, and M. Gopalakrishan. Speaker-independent recognition of spoken English letters. *IJCNN*, 2:45–51, 1990.

[5] R. Coultrip and R. Granger. Sparse random networks with LTP learning rules approximate Bayes classifiers via Parzen's method. *Neural Networks*, 7:463–476, 1994.

[6] R. Coultrip, R. Granger, and G. Lynch. A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks*, 5:47–54, 1992.

[7] A. Garzotto, B. Aleksandrovsky, and R. Granger. A neocortically-derived model of continuous contextual processing. 1996. (in preparation).

[8] M. Gluck and R. Granger. Computational models of the neural bases of learning and memory. *Annual Review of Neuroscience*, 16:667–706, 1993.

[9] R. Granger, J. Whitson, J. Larson, and G. Lynch. Non-hebbian properties of long-term potentiation enable high-capacity encoding of temporal sequences. *Proc. Nat'l. Acad. Sci.*, 91:10104–10108, 1994.

[10] K. Kilborn, R. Granger, and G. Lynch. Effects of LTP on response selectivity of simulated cortical neurons. *J. Cognitive Neuroscience*, 1996. (in press).

[11] J. Larson and G. Lynch. Theta pattern stimulation and the induction of LTP: the sequence in which synapses are stimulated determines the degree to which they potentiate. *Brain Research*, 489:49–58, 1989.

[12] J. Larson, D. Wong, and G. Lynch. Patterned stimulation at the theta frequency is optimal for induction of long-term potentiation. *Brain Research*, 386:347–350, 1986.

[13] R. Lyon. A computational model of filtering, detection and compression in the cochlea. *Proceedings of ICASSP*, 1982.

[14] J. Makhoul. Spectral linear prediction: properties and applications. *IEEE Transactions on ASSP*, 23:283–296, 1975.

[15] Y. Muthusamy, R. Cole, and M. Slaney. Speaker-independent vowel recognition: spectograms versus cochleograms. *Proceedings of ICASSP*, 1990.

[16] J. Whitson, B. Aleksandrovsky, E. Whelpley, G. Lynch, and R. Granger. Searching for a general cortical memory circuit: A computational model based upon cortical homologies stores, retrieves, and recognizes feature sequences. *J. Cognitive Neuroscience*, 1996. (submitted).