Exploiting Temporal Coherence in Speech for Data-Driven Feature Extraction

Michael A. Carlin[1,2,3] and Mounya Elhilali[1,3]
[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
[3]Dept. of Electrical and Computer Engineering
The Johns Hopkins University
Baltimore, MD 21218, USA
Email: {macarlin, mounya}@jhu.edu

***Abstract*—It is well known that speech sounds evolve at multiple timescales over the course of tens to hundreds of milliseconds. Such temporal modulations are crucial for speech perception and are believed to directly influence the underlying code for representing acoustic stimuli. The present work seeks to explicitly quantify this relationship using the principle of *temporal coherence*. Here we show that by constraining the outputs of model linear neurons to be highly correlated over timescales relevant to speech, we observe the emergence of neural response fields that are bandpass, localized, and reflective of the rich spectro-temporal structure present in speech. The emergent response fields also appear to share qualitative similarities those observed in auditory neurophysiology. Importantly, learning is accomplished using unlabeled speech data, and the emergent neural properties well-characterize the spectro-temporal statistics of the input. We analyze the characteristics and coverage of ensembles of learned response fields for a variety of timescales, and suggest uses of such a coherence learning framework for common speech tasks.**

I. Introduction

With no supervision, the auditory system is tasked with representing sound in such a way that we can separate competing sources, discriminate among a variety of acoustic classes, and, in general, process complex auditory scenes. However, the principles governing how the auditory system extracts useful information from the environment remain unclear. A general proposal is that the system uses knowledge of physical constraints of the acoustic soundscape and the statistics of environmental sounds to solve these complex tasks. This knowledge is believed to be reflected in the sensory processing and neural encoding of incoming sounds. The statistics of natural sounds are quite complex and particularly so for speech, the most important communications sound for humans. Well-known results from psychoacoustics [1] and more recently in physiology [2] have shown that speech sounds are processed at multiple concurrent timescales, from the segmental level ($\sim$20–50 ms) to suprasegmental ($\sim$150–300 ms) levels. Somehow, the neural code needs to capture these different dynamics while still yielding a stable perception of the environment.

Recent evidence from neurophysiology suggests that *sustained* neural firings by central auditory neurons form an important part of the coding strategy for representing the acoustic environment [3]. To what extent these responses capture the relevant temporal modulations of speech remains unclear. In this paper, we study a method to explicitly encode the notion of sustained neural responses over speech-relevant timescales using the principle of *temporal coherence* [4]. By enforcing the outputs of model neurons to be highly correlated over specific time intervals, we show the emergence of receptive fields that are bandpass, highly localized, and reflective of the spectro-temporal structure of the stimulus, much like those observed in physiological studies [5], [6]. We then analyze the spectro-temporal characteristics of the emergent ensembles at a variety of timescales relevant to speech. Finally, we note that learning is accomplished using unlabeled speech data, and the emergent neural properties appear to well-characterize the statistics of the input. Such a framework has implications for robust processing of speech signals and we suggest uses of the emergent ensembles for common speech tasks.

We begin by presenting background and previous work in Section II. This is followed in Section III by a description of a quantitative model of temporal coherence and an optimization procedure for enforcing temporal coherence using speech stimuli. In Section IV we present the main results of this work, followed by suggestions for applications to common speech tasks in Section V.

II. Background

It is widely believed that sensory representations over many modalities and timescales are shaped by the environment in which an organism operates. However, since sensory systems are often not given explicit supervision for learning how to represent and discriminate among competing stimuli, determining the underlying principles governing choice of a sensory code remains an open question. A general approach that has found success particularly in the visual and auditory domains involves quantifying the relationship between environmental statistics and sensory representation by design of a suitable cost function. One can then study how the solution to the objective function relates to known properties of the sensory modality [7]. In vision, for example, Olshausen and Field [8] demonstrated that learning to represent static natural images using a sparse code yields a set of localized and oriented

978-1-4244-9848-2/11$26.00©2011 IEEE

Gabor-like basis functions resembling the shapes of receptive fields observed in simple cells of the visual cortex. In audition, Lewicki [9] proposed that the auditory system uses what he termed an efficient code for representing natural sounds in the auditory periphery. There he demonstrated that modeling efficiency by enforcing statistical independence on the output of model cochlear filters in response to natural sounds yields an analysis filterbank with properties remarkably similar to those observed in auditory physiology.

While computational models of sensory representations at the cochlear level have began unraveling the principles guiding peripheral processing, the relationship between acoustic signals and representation in central auditory areas remains unclear. For instance, Hromadka *et al.* [10] presented evidence suggesting that auditory cortex is optimized to use a sparse code by showing that at any instant, the ensemble response to a variety of acoustic stimuli involved only a small percentage of neurons eliciting brief, transient firing rates. This finding appears to be at odds with results reported by Wang *et al.* [3] showing that while population responses may indeed show transient firing rates, subsets of neurons will exhibit strong, *sustained* responses when presented with preferred stimuli. Furthermore, by also noting a lack of stimulus synchrony for these persistent neurons, Wang *et al.* have argued that a sustained firing rate represents a transformation of the acoustic stimulus and not merely a preservation of the temporal dynamics of the input sound. It is precisely this idea that motivates the present work.

Assuming that the notion of *sustained* neural firing rates forms part of the underlying auditory code, it is then necessary to consider the timescales over which this behavior may relate to the temporal statistics of speech stimuli. Evidence from physiology [2], [11] and psychoacoustics [1] suggests that speech cues are decoded primarily based temporal modulations at the *segmental* level, ranging from ∼20–50 ms, as well as at the *suprasegmental* level, ranging from ∼150–300 ms. The segmental level would thus capture pitch and formant (place of articulation) cues whereas the suprasegmental level captures syllabic and prosodic cues. Indeed, systematic degradations of temporal modulations at both these timescales significantly impair the ability of listeners to comprehend speech [12], [13].

In this work, we seek to understand if a relationship exists between the temporal statistics of speech and the use of sustained neural responses as a coding strategy in the central auditory system. To probe this question, we explicitly quantify neural persistence using *temporal coherence*, which has previously been shown to relate the temporal statistics of natural image sequences to receptive field characteristics in primary visual cortex [4]. We present an adapted version of this model for processing acoustic stimuli, and explore the effect on the receptive fields of model auditory neurons when explicitly enforcing sustained responses on a neural ensemble.

III. Model

A typical assumption made when studying the mapping between an acoustic signal $s(t, f)$ and the firing rate $r(t)$ of a central auditory neuron is that the transformation is linear [5], i.e.,

$$r(t) = \sum_f \sum_m s(t-m, f)h(m, f) \tag{1}$$

where $h(t, f)$ is the *spectro-temporal receptive field* (STRF) of the neuron, represented as an LTI filter in time and frequency. For discrete-valued signals and filters, we can stack $s(t, f)$ and $h(t, f)$ columnwise to obtain vector representations $\mathbf{s}(t)$ and $\mathbf{h}$, respectively, and compactly write the firing rate as

$$r(t) = \mathbf{h}^T \mathbf{s}(t). \tag{2}$$

Furthermore, we can express the response $\mathbf{r}(t) = [r_1(t)\, r_2(t)\, \cdots\, r_K(t)]$ of an *ensemble* of K neurons by concatenating the STRFs in to a matrix $H := [\mathbf{h}_1\, \mathbf{h}_2\, \cdots\, \mathbf{h}_K]$ and writing

$$\mathbf{r}(t) = H^T \mathbf{s}(t). \tag{3}$$

To quantify the notion of a sustained firing rate over an interval $[t - \Delta T, t]$, we adapt the model of [4] and define *temporal coherence* as

$$J := \sum_{k=1}^{K} \sum_{\tau=1}^{\Delta T} \alpha_\tau E_t \left[r_k^2(t) r_k^2(t-\tau) \right], \tag{4}$$

where $E_t[\cdot]$ denotes expected value over time. Here the notion of coherence is quantified by correlation between signal energies over a *coherence interval* specified by ΔT. Hence, if the $r_k(t)$ vary smoothly over the coherence interval, as would be the case for a sustained neural response, we would expect J to be large. The weights α_τ are chosen to reflect the intuition that recent observations likely have more influence on the current output than those from the past; in this work the α_τ are set to be linear.

Thus, a statement about enforcing sustained responses over an interval $[t - \Delta T, t]$ corresponds to maximizing temporal coherence. We can therefore define the following optimization problem:

$$\underset{\mathbf{h}_1\, \mathbf{h}_2 \cdots \mathbf{h}_K}{\operatorname{argmax}} \; J \tag{5}$$

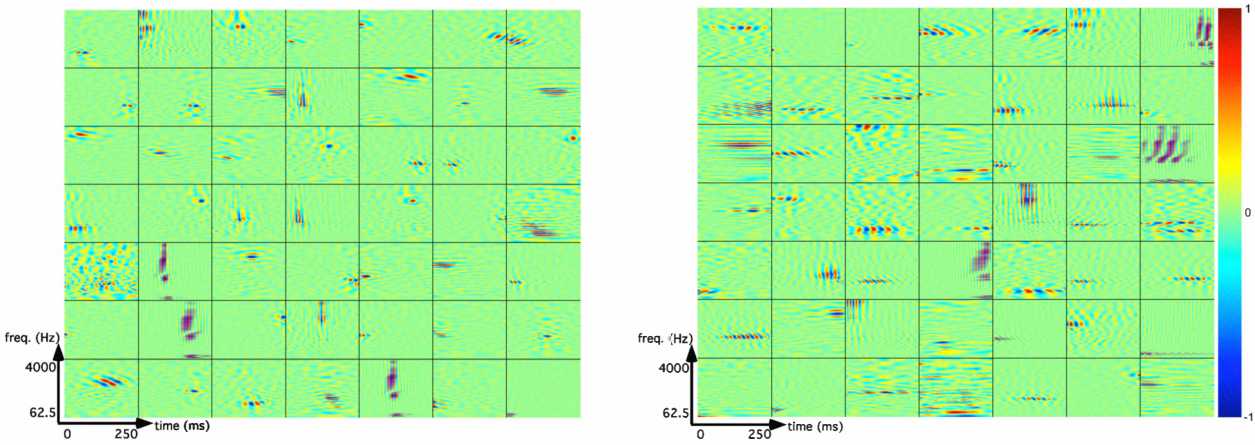
subject to

$$\begin{aligned} E_t[r_k^2(t)] &= 1, \qquad \forall\, k \\ E_t[r_k(t) r_l(t)] &= 0, \quad \forall k \neq l \end{aligned} \tag{6}$$

for $k, l \in \{1, 2, \cdots, K\}$. We impose these constraints to (1) bound the output of the model neurons and prevent the trivial solution $r_k(t) = 0$ and (2) minimize redundancy of the solution by requiring that the responses of different neurons be uncorrelated.

Optimization of the above nonlinear program is accomplished using a variant of the *symmetric orthogonal projection algorithm* presented in detail in [4]. In essence, learning the ensemble of STRFs H that maximize temporal coherence is accomplished via gradient ascent on Eq. 4 with a suitable projection of the sequence of updates $H^{(n)}$, $n = 1, 2, \cdots, N$, so as to satisfy the constraints in Eq. 6. We refer the reader to [4] for basic implementation details. To accommodate learning on potentially large datasets, we updated the procedure to use

978-1-4244-9848-2/11$26.00©2011 IEEE



(a) $\Delta T = 10$ ms (b) $\Delta T = 250$ ms

Fig. 1: Examples of emergent STRFs for coherence intervals of 10 and 250 ms.

stochastic gradient ascent using smaller batches of training data.

IV. Results

A. Emergence of localized spectro-temporal receptive fields

For the following experiments, the input speech stimuli were obtained from the TIMIT speech corpus with waveforms sampled at 8 kHz. For each utterance, an auditory spectrogram was computed yielding a representation of the acoustic signal in time and (logarithmic) frequency [14]. The auditory spectrograms were then segmented and vectorized as described above. Each segment covered 250 ms in time and six octaves in frequency (62.5–4000 Hz), and a segment was extracted every 5 ms from the original spectrograms. In total, 25k (columnwise) samples generated from equal amounts of male and female utterances were used as training data. A set of $K = 400$ STRFs were initialized at random and optimized via the procedure discussed in the previous section. Empirically, we found that using 5k sample batches with 30 iterations through the complete training set yielded suitable results. We considered coherence intervals in the segmental and suprasegmental range (10–500 ms), as well as an extended interval for ΔT at 2.5 sec to observe any changes for long integration windows.

Examples of the resulting STRFs for $\Delta T = \{10, 250\}$ ms are shown in Fig. 1. Each patch shown represents the tuning of an STRF over 250 ms in time and six octaves in frequency (62.5–4000 Hz). The red and blue colors indicate that the presence or absence, respectively, of energy in a particular region causes a strong neural response. One will readily observe a variety of STRF tunings: sensitivity to pitch and harmonicity; sensitivity to temporally fast and spectrally broad sounds as with plosives and fricatives; selectivity to highly localized and narrowband regions of energy; and directionally sensitive tuning to spectro-temporal transitions as observed with formants, for example. These general classes of STRFs are also qualitatively similar to observations made in neurophysiology [5], [6]. In addition to the diversity in shapes of the emergent STRFs, we also note the diversity of phases of some of the basic shapes, indicating broad ensemble coverage of time and frequency.

B. Analysis of Emergent Ensembles

To examine how the structure of the emergent ensembles changed with choice of coherence interval, we varied ΔT to enforce coherence at those timescales important for speech, i.e., segmental scales on the order of tens of milliseconds and suprasegmental scales on the order of hundreds of milliseconds. In the context of a linear model, it was anticipated that to maintain persistence for increasing ΔT, the STRFs must account for more energy in time and frequency and consequently may necessarily become broader in both dimensions. The first observation we made to this effect was that temporal bandwidths of the STRFs in each ensemble tended to increase with increasing ΔT. This was quantified by first performing a least-squares fit of a Gaussian envelope to each STRF in a given ensemble, summing the envelope along the spectral axis to yield a smooth temporal profile, and calculating the 10-dB excitatory bandwidth of the temporal profile. The distribution of temporal bandwidths is shown in Figure 2. As observed, increasing the coherence interval induces a corresponding increase in temporal bandwidth. No significant changes in spectral bandwidth were observed.

To further quantify apparent structural variations we observed with varying ΔT, we manually labeled each STRF according to one of six broad classes we determined to vary across ensembles. Examples of each of these classes are shown in Fig. 3 and included the following types: localized (*local*), spectral (*spec*), directional (*dir*), temporal (*temp*), noisy (*noise*), and complex (*cplx*).

Given the assignments, we then sought to determine if there was an "optimal" ΔT for which diversity across each of the classes was somehow balanced to maximize sensitivity to a variety of acoustic classes. To address this question,

978-1-4244-9848-2/11$26.00©2011 IEEE

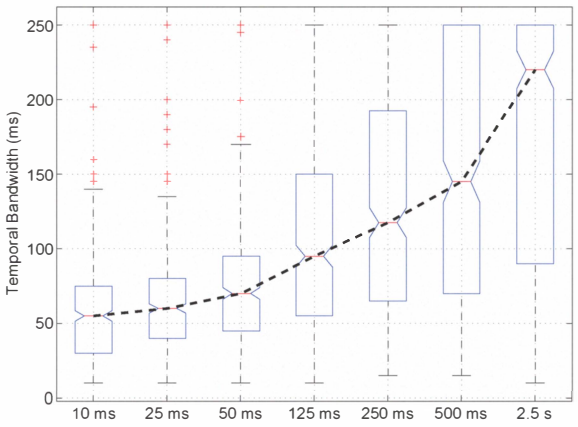


Fig. 2: Distributions of STRF temporal bandwidths with increasing ΔT.

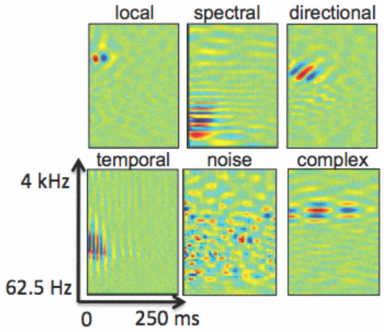


Fig. 3: Examples of broad classes of STRFs that varied with different choices of ΔT.

we considered how class membership within an ensemble changed with ΔT. Shown in Fig. 4 are the percent of STRFs for a particular ΔT labeled according to the broad classes described above. There are a few basic observations. First, the shorter coherence intervals ($\Delta T = 10, 25, 50$) ms tend to be dominated by highly localized STRFs, perhaps reflecting sensitivity to more segmental cues. Second, we observe that the percent of noisy STRFs is largest for $\Delta T = 2.5$ sec, indicating an overall loss of structure for longer coherence intervals. Third, there is a pooling of local maxima and minima for the spectral, complex, and directional classes for ΔT between 50–500 ms, suggesting a potential tradeoff between sensitivity to segmental and suprasegmental cues in this range. Finally, we note that there does not appear to be much variation with regard to the temporal STRFs across ΔT.

Since the previous two analyses focused on the structural variations of the emergent ensembles, we finally sought to compare variations in the statistics of the *outputs* of the neural ensembles by applying speech samples to their input once they had been trained according to the coherence objective. We used a small held-out speech set, also from the TIMIT corpus, comprising spectro-temporal segments from 50 male and 50 female utterances. We used the speech segments as input to each ensemble and considered the question of how

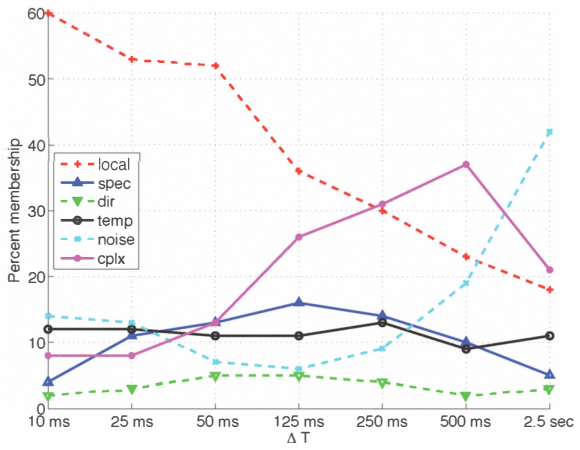


Fig. 4: Percent membership with respect to each of the broad classes as described in the text with increasing ΔT.

long a neuron tended to exhibit a sustained response once it became active. For a given ΔT, the k'th neuron was defined to be "active" when the absolute value of its firing rate $|r_k(t)|$ exceeded a fixed threshold (chosen to be $+1$ std. dev. of the k'th STRF output), and we collected duration statistics for all STRFs. The $K = 400$ STRFs in each ensemble were sorted by median activation time, and in Fig. 5 we report the distribution of median activations for the top 5% "most persistent" neurons for varying ΔT. Indeed, we observe an increase in duration of sustained responses of a subset of each ensemble with increasing ΔT, with a peak at 250 ms in the suprasegmental timescale.

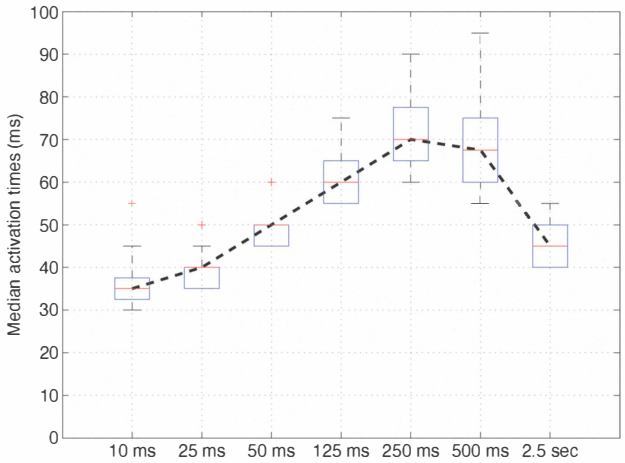


Fig. 5: Median activation times of the top 5% "most persistent" STRFs for increasing ΔT.

V. Discussion and Conclusions

As demonstrated, maximizing temporal coherence using unlabeled speech stimuli yields diverse ensembles of STRFs which appear to capture the relevant structure of speech sounds both in time and frequency. We also note that by having

978-1-4244-9848-2/11$26.00©2011 IEEE

directly enforced sustained responses in an ensemble of model neurons, the trends observed across ΔT in both the broad STRF assignments as well as the activations of the "most persistent" neurons suggest that the proposed model does indeed capture the relevant timescales for processing speech sounds. Current work is focusing on procedures for automated clustering of the STRF ensembles to eliminate any potential bias in the manual annotations, but the reported assignments capture in spirit the trends we observed when comparing emergent ensembles across ΔT.

An immediate benefit for speech systems may be gained by the basic observation that often the STRFs are highly localized, especially for short ΔT, which may prove useful for robust signal detection in noise. Furthermore, as observed in Fig. 5, the tendency of a neuron to remain active with increasing ΔT may prove particularly beneficial when the input speech is subjected to temporal or channel distortions. Finally, as the STRFs discussed in this paper are learned entirely on unlabeled data with a balance between male and female read speech, it remains to be seen how the emergent ensembles may be biased to particular genders, speakers, speaking styles, and acoustic channels, knowledge of which in most cases is beneficial for common speech tasks.

ACKNOWLEDGMENTS

This research was supported by a graduate fellowship from the Human Language Technology Center of Excellence, NSF CAREER grant IIS-0846112, AFOSR grant FA9550-09-1-0234 and NIH grant 5R01AG036424.

REFERENCES

[1] S. Rosen, "Temporal information in speech: acoustic, auditory, and linguistic aspects," *Phil. Trans. R. Soc. Lond. B*, vol. 336, pp. 367–373, 1992.

[2] D. Poeppel, "The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'," *Speech Comm.*, vol. 41, pp. 245–255, 2003.

[3] X. Wang, T. Lu, R. K. Snider, and L. Liang, "Sustained firing in auditory cortex evoked by preferred stimuli," *Nature*, vol. 435, pp. 341–346, 2005.

[4] J. Hurri and A. Hyvarinen, "Simple-cell-like receptive fields maximize temporal coherence in natural video," *Neural Comp.*, vol. 15, pp. 663–691, 2003.

[5] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophys.*, vol. 85, pp. 1220–1234, 2001.

[6] F. E. Theunissen, K. Sen, and A. J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *J. Neurosci.*, vol. 20, no. 6, pp. 2315–2331, 2000.

[7] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, 2001.

[8] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[9] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neurosci.*, vol. 5, no. 4, pp. 356–363, 2002.

[10] T. Hromadka, M. R. DeWeese, and A. M. Zador, "Sparse representation of sounds in the unanesthetized auditory cortex," *PLoS Bio.*, vol. 6, no. 1, p. e16, 2008.

[11] G. Hickok and D. Poeppel, "The cortical organization of speech processin," *Nature Neurosci.*, vol. 8, pp. 393–402, 2007.

[12] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[13] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comp. Bio.*, vol. 5, no. 3, p. e1000302, 2009.

[14] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.

978-1-4244-9848-2/11$26.00©2011 IEEE