Robust Phoneme Recognition Based on Biomimetic Speech Contours

Michael A. Carlin, Kailash Patil, Sridhar Krishna Nemala, and Mounya Elhilali

Dept. of Electrical and Computer Engineering & Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD USA
{macarlin,kailash,nemala,mounya}@jhu.edu

Abstract

It has been previously suggested that ensembles of central auditory neurons optimize a sustained firing criterion as part of the underlying neural code for representing sound. Moreover, computational studies have shown that optimizing such a criterion yields ensembles of spectro-temporal receptive fields akin to those observed in physiological studies. In this study, we show that these emergent receptive fields contour the high-energy modulations in speech, defining a boundary that distinguishes between noise-robust and easily corrupted modulations in speech-plus-noise mixtures. A simple 2D filter thus derived is shown to improve upon the performance of state-of-the-art phoneme recognition systems under both additive noise conditions and reverberation by 5.9% absolute on average.

Index Terms: robust feature extraction, bio-inspired features, sustained neural firings

1. Introduction

A critical component of automatic speech recognition systems is the choice of features for representing the acoustic signal. Such features should not only be easy to compute but also exhibit some degree of noise robustness to inevitable degradations to the acoustic signal when used in real environments. However, it is often the case that the performance of sophisticated feature extraction schemes, while demonstrating state-of-the-art performance in clean acoustic conditions, quickly degrades in the presence of additive noise or reverberation.

Motivated by the robustness of the mammalian auditory system in degraded acoustic environments, it is believed that observations from behavioral and neurophysiological studies can inform processing schemes for automated sound processing systems. For instance, it is widely believed that "slow" spectro-temporal modulations in speech carry information in a robust manner in degraded acoustic environments [1, 2, 3]. Application of this principle has recently been shown by Nemala *et al.* to identify those spectro-temporal modulations that yield noise-robust features when corrupted by a variety of additive noise conditions [4].

Additionally, studies of the basis of sound representation in central auditory areas suggest that *sustained neural responses* form part of the code underlying the perceptual stability of auditory objects [5, 6, 7]. Indeed, a computational model considered by Carlin and Elhilali that enforces sustained responses yields ensembles of spectro-temporal receptive fields (STRFs) akin to those measured in physiological studies [8, 9]. Analysis of the modulation profiles of the emergent STRFs suggests that spectro-temporal modulation contours serve to distinguish between noise-robust and easily corrupted modulations in speech-plus-noise mixtures. Importantly, these results complement the findings of Nemala *et al.*, and in this paper we elaborate on this relationship.

Here we describe how the sustained firing principle can be used to derive a data-driven 2D spectro-temporal modulation filter for preprocessing auditory spectrograms for noise-robust feature extraction. In a phoneme recognition task, we demonstrate that use of these filtered spectrograms outperform state-of-the-art mean-variance ARMA (MVA) features in both additive noise and reverberant conditions.

2. Methods

2.1. Spectro-temporal receptive fields

To characterize the relationship between a stimulus and its corresponding neural response we use the spectro-temporal receptive field (STRF) [10]. An STRF models the linear transformation of a time-varying spectro-temporal input to an instantaneous firing rate, i.e.,

$$r(t) = \int\int h(\tau, f)s(t-\tau, f)d\tau df \tag{1}$$

where $h(t, f)$ is an LTI filter that defines the STRF and $s(t, f)$ is a spectro-temporal stimulus. For discrete-time signals and assuming that $h(t, f)$ has a finite impulse response, we can express Eq. 1 compactly in vector notation as

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

where $\mathbf{s}(t), \mathbf{h} \in \mathbb{R}^d$ are vectors denoting the (column-wise stacked) stimulus and filter, respectively [11]. Furthermore, to express the response $\mathbf{r}(t) = [r_1(t)\, r_2(t) \cdots r_K(t)]^T \in \mathbb{R}^K$ of an *ensemble* of K neurons, we concatenate the STRFs into a matrix $H := [\mathbf{h}_1\, \mathbf{h}_2 \cdots \mathbf{h}_K] \in \mathbb{R}^{d\times K}$, which allows us to write the *ensemble* response as $\mathbf{r}(t) = H^T\mathbf{s}(t)$.

2.2. Optimizing a Sustained Firing Criterion

A sustained response can be understood as one whose firing rate changes relatively slowly and is thus highly *correlated* over time. Here we are interested in the characteristics of ensembles of model STRFs H that promote sustained responses over a specified time interval $[t-\Delta T, t]$. To quantify this principle, we adapt the model of Hurri and Hyvarinen [12] and define the following objective function:

$$J_{sus}(H) := \sum_{k=1}^{K}\int_{\Delta T}\alpha_\tau\left\langle r_k^2(t)r_k^2(t-\tau)\right\rangle_t d\tau, \tag{3}$$

This work was partially supported by a graduate fellowship from the Human Language Technology Center of Excellence, and grants IIS-0846112 (NSF), FA9550-09-1-0234 (AFOSR), 1R01AG036424 (NIH) and N000141010278 (ONR).

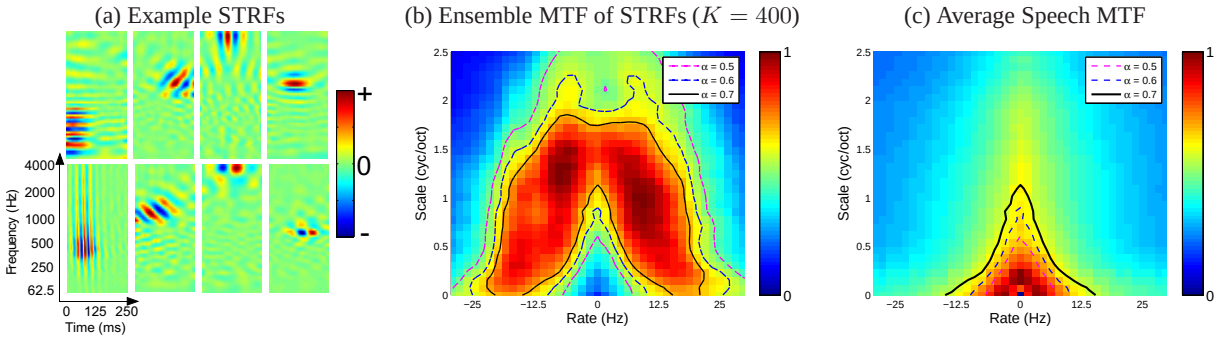

Figure 1: *(a) Examples of emergent STRFs learned by optimizing the sustained firing criterion for $\Delta T = 125$ ms, (b) the corresponding STRF ensemble MTF (eMTF), and (c) an estimate of the average speech MTF. In panels (b) and (c) we superimpose normalized isoline contours derived from the eMTF at various α-levels. For display purposes, the MTF in (c) is compressed by a factor of $1/3$.*

where $\langle\cdot\rangle_t$ denotes time average. Observe that $J_{sus}(H)$ represents the sum of correlations between signal energies of the k'th neuron over a time interval defined by ΔT across an ensemble of K neurons. If a neuron yields a sustained response, then each of the $r_k(t)$ vary smoothly over the specified interval and we expect $J_{sus}(H)$ to be large. Moreover, choice of ΔT allows us to directly explore the effect of different timescales on the ensembles H that optimize Eq. 3. Finally, the weights α_τ are chosen to reflect the intuition that recent activity of a neuron likely has more influence on the current output than the past; in this work the α_τ are set to be linearly decaying.

Since the ensemble H is not specified *a priori*, the goal is to vary the shapes of the STRFs so as to *maximize* sustained firing rates according to the objective function defined in Eq. 3, subject to constraints that *bound* the responses and *minimize redundancy* in the learned ensemble. Such constraints can be satisfied by enforcing the responses have unit variance and be mutually uncorrelated [8, 12]. Thus, we wish to solve the following optimization problem:

$$\arg\max_{H} \; J_{sus}(H) \;\text{ subject to }\; \left\langle r_j(t) r_k(t) \right\rangle_t = \delta_{jk}, \qquad (4)$$

for $j, k = 1, 2, \cdots, K$ and where δ_{jk} is the Kroenecker delta function. For brevity's sake, we omit a detailed description of the optimization procedure[1]. However, it suffices to say that the matrix of STRFs H is updated via projected gradient ascent whereby a projection $\mathcal{P}(\cdot) : \mathbb{R}^{d\times K} \rightarrow \mathbb{R}^{d\times K}$ is applied to each gradient update $H^{(i)}$ so that the required response constraints are satisfied; the interested reader is directed to [12] for more details.

To learn STRFs that optimize the sustained firing criterion, we used auditory spectrograms [13] computed from approx. three minutes of speech from the TIMIT `train` corpus, using an equal proportion of male and female speakers. The tonotopic axis was sampled using 10 channels/octave over 6 octaves at a frame rate of 5 ms. We extracted 250 ms spectro-temporal segments once every 5 ms. Each segment was stacked column-wise into a vector $\mathbf{s}(t) \in \mathbb{R}^d$ where $d = 3000$ (i.e., 50 vectors/segment $\times$ 60 channels), yielding a total of $\sim$30k spectro-temporal input vectors. An ensemble of STRFs H was initialized at random and varied so as to solve the problem posed in Eq. 4 above. Examples of STRFs learned using the above procedure for $\Delta T = 125$ ms are shown in Fig. 1(a). As observed, the STRFs exhibit sensitivity to a variety of localized, spectral, temporal, and joint spectro-temporal events in the stimulus.

[1] A detailed description of this procedure along with full analysis on a broader set of natural sounds will appear in a future paper.

2.3. Modulation Analysis of the Emergent STRFs

A useful characterization of the spectro-temporal modulation sensitivity of an STRF is made by considering its *modulation transfer function* (MTF). The MTF is simply the magnitude of the 2D Fourier transform of a given STRF and describes the joint distribution of sensitivity to temporal modulations (*rate*, in Hz) and spectral modulations (*scale*, in cyc/oct). Furthermore, by averaging the MTFs obtained from each STRF, we obtain an *ensemble MTF* (eMTF) that characterizes the average spectro-temporal modulation sensitivity of the given ensemble [14]. The eMTF can then be used to relate the average modulation tuning of an ensemble to the modulations present in the stimulus.

Shown in Fig. 1(b) is the normalized eMTF for an ensemble of $K = 400$ STRFs, again for $\Delta T = 125$ ms. Interestingly, the eMTF shows that the emergent STRF ensemble has little-to-no sensitivity to "slow" modulations (i.e., no energy close to the origin), exhibiting instead a distinct "contouring" effect for rates between approx. $\pm$ 15 Hz and scales between 0 and 2 cyc/oct. It is known, however, that speech has an abundance of modulation energy in these modulation ranges [2], and indeed this is observed when we compute the average MTF of the speech stimulus (Fig. 1(c)).

To compare the extent to which the modulation energy of the STRFs contours the modulations of the speech stimulus, we computed normalized isoline contours at the α level (Fig. 1(b)), and considered those portions of the contours closest to the origin (Fig. 1(c)). Indeed, when superimposed on the speech MTF, we observe that the contours form a tight boundary around those rates and scales where most of the speech modulation energy is concentrated. This is an especially interesting observation given the recent results of Nemala *et al.* [4], who have demonstrated that auditory spectrograms bandpass filtered to contain only "slow" rates and scales in this region yield noise-robust features for automatic speech recognition. To complement these results, we hypothesize that the observed contouring effect due to the eMTF serves to define the band edges of a 2D bandpass modulation filter in a *data-driven* fashion. We describe next how this principle is used to design such a filter.

2.4. 2D Spectro-Temporal Modulation Filtering

The 2D filters we consider are designed in the modulation domain using a given contour C at the α level. We set the magnitude response of the filter at rates and scales inside the contour to unity. We then set the roll-off of the filter to be exponential

as

$$M_1(\omega,\Omega) = \exp\left\{-\left(\frac{(\omega-\omega_c)^2}{\omega_r} + \frac{(\Omega-\Omega_c)^2}{\Omega_s}\right)\right\} \quad (5)$$

where (ω_c, Ω_c) is the point from C that is closest to the point (ω, Ω) being considered. Here, ω_r and Ω_s are the roll-off parameters along the rate and scale axis, respectively. To remove temporal modulations near 0 Hz, we define a wedge function as

$$W(\omega) = \begin{cases} \sin\left(\frac{\pi\omega}{2\omega_W}\right) & |\omega| < \omega_W \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

where ω_W is the wedge roll-off along the rate axis [3]. Thus, we obtain the desired 2D filter as $M(\omega,\Omega) = M_1(\omega,\Omega) \cdot W(\omega)$. A given auditory spectrogram is filtered by first transforming to the modulation domain via the 2D Fourier transform, and the magnitude is multiplied with the filter $M(\omega,\Omega)$. Finally, we perform the inverse 2D Fourier Transform, keeping the real part only, to obtain the filtered auditory spectrogram.

3. Experiments and Results

3.1. Corpora and Recognizer Setup

Hand-labeled data from the TIMIT corpus was used to train a speaker-independent phoneme recognition system using the Hybrid Multi-layered Perceptron / Hidden Markov Model (MLP/HMM) setup [15]. 3696 utterances were used for training out of which 8% were used as cross validation data. A separate set of 1344 utterances were used for testing. The 61 phoneme labels in the TIMIT corpus were converted to a standard set of 39 labels [16].

A multi-layered perceptron (MLP) was trained discriminatively to estimate the posterior probabilities of the phoneme classes given an input feature vector. The MLP had a hidden layer with 1500 nodes with a sigmoid non-linearity. The output layer consisted of 40 nodes (with a softmax non-linearity) corresponding to the 39 phonemes and an additional garbage class. A second MLP was then trained to include a temporal context of 23 frames (11 frames before and after the current frame) and helped to enhance the posterior probability estimates. The second MLP had the same hidden layer and output layer structure as the first [17].

The HMM system consisted of a three-state feed-forward HMM for each phoneme, with equal probability of transition to itself or the next state. The posterior probabilities were divided by the relative counts of each phoneme and were used as the emission probabilities for the HMM. Finally, the phoneme sequence was decoded using the standard Viterbi algorithm. This decoded sequence of phonemes was compared to the hand-labeled sequence, with recognition rate determined by the number of insertions, deletions, and substitutions.

To assess the noise-robustness of the proposed features, we tested the system under various mismatched conditions. For this we corrupted the test set with additive noise and reverberation. Five types of additive noises from the NOISEX92 corpus [18] were added to the test data at various SNRs from 0–20 dB (at steps of 5 dB) using the FaNT tool [19]. The noises considered were speech babble (Babble), fighter jet cockpit (F16), factory floor (Factory1), military tank (Tank), and automobile interior (Volvo). For reverberation, we synthesized artificial room responses at five different reverberation time constants (RT_{60}) from 100–500 ms in steps of 100 ms. These responses were generated by convolving Gaussian white noise with an exponentially decaying envelope.

Table 1: Phoneme recognition rate (as %) for utterances corrupted by additive noise (higher is better).

Noise Type	SNR (in dB)	Feature	
		MFCC+MVA	2D Filtered
Clean	∞	68.2	69.6
Babble	20	56.6	63.8
	15	49.6	57.7
	10	40.7	47.8
	5	29.8	34.6
	0	19.6	21.8
	Average	39.3	45.1
F16	20	57.1	62.4
	15	50.8	56.5
	10	43.3	47.4
	5	34.6	37.2
	0	27.0	27.2
	Average	42.6	46.1
Factory1	20	55.8	61.6
	15	48.5	55.1
	10	39.5	46.2
	5	30.2	35.6
	0	21.2	25.9
	Average	39.0	44.9
Tank	20	57.8	67.1
	15	54.5	64.7
	10	50.7	60.3
	5	46.4	54.4
	0	41.4	46.5
	Average	50.1	58.6
Volvo	20	63.6	69.6
	15	62.0	69.3
	10	60.2	68.6
	5	58.1	67.2
	0	54.8	64.7
	Average	59.7	67.9

3.2. Proposed and Baseline Features

For the proposed features, the auditory spectrogram of each utterance was calculated at a spectral resolution of 24 channels per octave over 5.3 octaves (128 channels in total) at a frame rate of 100 frames/second. We used the contour derived for $\alpha = 0.7$, and 2D filtering (as described in Sec. 2.4) was applied with $\omega_r = 1, \Omega_s = 0.12$, and $\omega_W = 1.25$. These constants were empirically determined to maximize performance on the cross validation data set. After applying the 2D filter, we appended first-, second-, and third-order dynamic features, yielding a 512-dimensional input feature vector (i.e., 128×4).

We compared the proposed features with state-of-the-art noise robust features based on MVA processing of MFCC features [20]. These features were obtained by first extracting a standard set of 13-dimensional MFCCs including their first-, second-, and third-order temporal derivatives. Next, cepstral mean subtraction and variance normalization was applied, and the temporal trajectory of each feature dimension was filtered in a RASTA-like manner, further enhancing noise robustness [21]. Finally, a nine-frame context was appended, resulting in a 468-dimensional feature vector (i.e., $13\times4\times9$).

3.3. Results

Shown in Table 1 are phoneme recognition results for test utterances corrupted by additive noise at a variety of SNRs. It is immediately clear that for clean as well as for all noise types and noise levels the proposed features outperform the baseline MFCC+MVA features, with an overall average absolute gain

Table 2: Phoneme recognition rate (as %) for utterances corrupted by artificial reverberation (higher is better).

Reverb. time (RT_{60})	Feature	
	MFCC+MVA	2D Filtered
100 ms	50.1	53.4
200 ms	37.3	40.6
300 ms	30.5	34.3
400 ms	27.1	30.9
500 ms	24.6	28.3
Average	33.9	37.5

of 6.4% for the noise cases. This improvement in performance even at 0 dB SNR suggests that the 2D filter is indeed able to capture the high energy regions of speech and discard the noise regions effectively.

Shown next in Table 2 are phoneme recognition results for test utterances corrupted by artificial reverberation. Again, in all cases, we observe that the proposed features outperform the baseline, with an average absolute gain of 3.6%. This further validates the robustness of the filter in capturing the high energy speech regions.

4. Discussion and Conclusions

We have demonstrated that by optimizing a neurophysiologically plausible sustained firing objective, we observe the emergence of an ensemble of STRFs that collectively define a tight boundary for speech in the modulation domain. By isolating spectro-temporal contours from the emergent ensemble MTF, we have described a framework for designing a 2D spectro-temporal filter for preprocessing spectrograms for noise-robust feature extraction. Moreover, the proposed features outperform state-of-the-art MVA-processed MFCCs both in clean conditions and in all additive noise and reverberation scenarios considered here.

While we could have derived the filter contours directly from the speech MTF, we consider the question of the information content of spectro-temporal modulations from an alternative but complementary perspective. In particular, the spirit of the work of Nemala *et al.* was to focus resources on subsets of rates and scales that were somehow "linguistically important" and presumably carried the message-bearing components of speech. This was achieved by choosing modulation filter parameters that reflected this intuition in the joint spectro-temporal modulation domain, and is indeed consistent with the RASTA filtering framework of Hermansky and Morgan [21].

It is therefore noteworthy that the sustained firing objective function and associated constraints arrive at a similar notion of data-driven filter design. In this work, rather than designing the shape of the modulation filter by hand, we arrived at a noise-robust representation for speech by considering more generally the form of a neural coding strategy used in central auditory areas. Additionally, the emergent neural ensemble, while implicitly capturing the extent of the slow spectro-temporal modulations in the stimulus, primarily exhibits sensitivity to fast modulations relatively far from the origin. Such a distribution may reflect more generally a form of unsupervised learning that discriminates among the various classes of sounds present in speech [22]. Future work is needed to further elucidate the relationship between the form of the objective function and constraints, the modulation spectra of the emergent STRFs, and the distribution of the speech MTF.

5. References

[1] 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.

[2] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *J. Acoust. Soc. Am.*, vol. 114, no. 6, pp. 3394–3411, 2003.

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

[4] S. K. Nemala, K. Patil, and M. Elhilali, "Multistream bandpass modulation features for robust speech recognition," in *Interspeech*, 2011.

[5] 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.

[6] X. Wang, "Neural coding strategies in auditory cortex," *Hearing Research*, vol. 229, pp. 81–93, 2007.

[7] J. C. Middlebrooks, "Auditory cortex cheers the overture and listens through the finale," *Nature Neurosci.*, vol. 8, no. 7, pp. 851–852, 2005.

[8] M. A. Carlin and M. Elhilali, "Exploiting temporal coherence in speech for data-driven feature extraction," in *IEEE Conf. on Information Sciences and Systems (CISS)*, 2011.

[9] ——, "Sustained firing of model cortical neurons yields richly structured spectro-temporal receptive fields," in *Abstracts of the Thirty-fifth ARO Mid-Winter meeting , Volume 35, Mt. Royal, NJ: Association of Research Otolaryngologists*, 2012.

[10] 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.

[11] F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant, "Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli," *Neuron: Computation in Neural Systems*, vol. 12, pp. 289–316, 2001.

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

[13] 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.

[14] L. M. Miller, M. A. Escabi, H. L. Read, and C. E. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," *J. Neurophys.*, vol. 87, pp. 516–527, 2002.

[15] H. Bourlard and N. Morgan, *Connectionist speech recognition: A hybrid approach.* Kluwer Academic, Dordrecht, 1994.

[16] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 37, pp. 1641–1648, 1989.

[17] J. Pinto, S. Garimella, M. Magimai-Doss, H. Hermansky, and H. Bourlard, "Analyzing mlp-based hierarchical phoneme posterior probability estimator," *IEEE Trans. Speech and Audio Process*, vol. 19, pp. 225–241, 2011.

[18] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," Speech Research Unit, Defense Research Agency, Malvern, U.K., Tech. Rep., 1992.

[19] H. Hirsch. (2005) Fant: Filtering and noise adding tool. Date last viewed 04/01/2012. [Online]. Available: http://dnt.kr.hsnr.de/download.html

[20] C. Chen and J. Bilmes, "Mva processing of speech features," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 257–270, 2007.

[21] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 2, pp. 382–395, 1994.

[22] S. M. N. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," *Nature Neurosci.*, vol. 8, no. 10, pp. 1371–1379, 2005.