HHS Public Access

Author manuscript

IEEE/ACM Trans Audio Speech Lang Process. Author manuscript; available in PMC 2018 June 12.

Published in final edited form as:

IEEE/ACM Trans Audio Speech Lang Process. 2015 December ; 23(12): 2422–2433. doi:10.1109/TASLP.2015.2481179.

A Framework for Speech Activity Detection Using Adaptive Auditory Receptive Fields

Michael A. Carlin [Member, IEEE] and **Mounya Elhilali [Member, IEEE]**

Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA

Abstract

One of the hallmarks of sound processing in the brain is the ability of the nervous system to adapt to changing behavioral demands and surrounding soundscapes. It can dynamically shift sensory and cognitive resources to focus on relevant sounds. Neurophysiological studies indicate that this ability is supported by adaptively retuning the shapes of cortical spectro-temporal receptive fields (STRFs) to enhance features of target sounds while suppressing those of task-irrelevant distractors. Because an important component of human communication is the ability of a listener to dynamically track speech in noisy environments, the solution obtained by auditory neurophysiology implies a useful adaptation strategy for speech activity detection (SAD). SAD is an important first step in a number of automated speech processing systems, and performance is often reduced in highly noisy environments. In this paper, we describe how task-driven adaptation is induced in an ensemble of neurophysiological STRFs, and show how speech-adapted STRFs reorient themselves to enhance spectro-temporal modulations of speech while suppressing those associated with a variety of nonspeech sounds. We then show how an adapted ensemble of STRFs can better detect speech in unseen noisy environments compared to an unadapted ensemble and a noise-robust baseline. Finally, we use a stimulus reconstruction task to demonstrate how the adapted STRF ensemble better captures the spectrotemporal modulations of attended speech in clean and noisy conditions. Our results suggest that a biologically plausible adaptation framework can be applied to speech processing systems to dynamically adapt feature representations for improving noise robustness.

Index Terms

Adaptive filtering; neural plasticity; spectro-temporal receptive fields; speech activity detection (SAD); stimulus reconstruction

I. Introduction

Current sound technologies borrow numerous biomimetic mechanisms widely observed in the brain in order to augment their robust sound processing. However, they remain mostly passive systems constrained to scenarios and conditions for which they were trained. Importantly, they fail to take advantage of adaptive capabilities that underlie the brain's ability to intelligently alter its response to changing tasks and listening environments. For

example, we can easily follow the first flute solo among the cacophony of an orchestra or keep track of a friend's voice at a noisy cocktail party. Considerable effort has been focused on studying the nature of this adaptive processing at the perceptual and neurophysiological levels in humans as well as animal models [1]–[5]. Moreover, a better understanding of the mechanisms by which the auditory system defines and selectively enhances the acoustic foreground while minimizing the impact of a noisy background has significant implications for signal processing strategies in noisy environments.

At the neural level, manifestations of adaptive processing are intricate, operating at multiple processing scales from subcellular all the way through network levels [6], [7]. For individual neurons, changing behavioral tasks can induce changes in a neuron's spectro-temporal receptive field (STRF) [8], [9]. The STRF is a two-dimensional kernel in time and frequency that summarizes the linear processing characteristics of a neuron. It can be thought of as a filter that operates on incoming acoustic inputs in order to extract specific features from sounds. In mammalian primary auditory cortex, STRFs exhibit detailed sensitivities to a broad range of acoustic features and are particularly selective of spectro-temporal energy modulations that characterize slow changes in temporal envelopes and spectral profiles of natural sounds [10]–[14].

Beyond their inherent tuning to specific acoustic modulations in the signal, cortical neurons can dynamically adapt their filtering properties towards relevant sounds in a task-driven manner. When cognitive resources are directed towards a sound of interest, cognitive feedback is believed to induce STRF plasticity, whereby cortical filters adapt from some "default" tuning to *task-optimal* shapes in order to enhance neural responses of task-relevant features while suppressing those of distractors. Such plasticity patterns have been observed in a number of neurophysiological studies involving simple tonal stimuli [15]–[17] as well as stimuli characterized by complex spectro-temporal dynamics [18]–[20]. The overall adaptation patterns reflect that of a *contrast-matched filter*, and it has been hypothesized that such changes serve to improve discriminability between the acoustic foreground and background [17]. Importantly, similar effects have been observed in other sensory modalities [21]–[23], suggesting that such discriminative changes in receptive field shape represent a general strategy used by sensory cortex to highlight task-relevant stimuli. The present work aims to understand the relevance of task-driven adaptation of filter tuning properties in speech processing tasks.

Human listeners are especially adept at tracking speech sounds in very noisy environments [24]. This ability engages a number of complex sensory and cognitive processes, most notably adaptive neural plasticity which allows the brain to hone in on conversations of interest. Speech signals are well characterized by their spectro-temporal modulation content, and so the Fourier domain is a natural space for exploring adaptive feature extraction strategies for speech signals. This is because a number of speech features can be expressed jointly in the spectro-temporal modulation domain, including voicing state, phoneme identity, formant trajectories, and syllable rate [25]–[28]. Furthermore, sounds having considerable overlap in time-frequency may in fact be disjoint in the modulation domain, leading to methods for signal denoising and enhancement [29]. Finally, modulation-domain adaptation has connections to a general form of object-based cognitive feedback.

Specifically, Fourier-based analysis facilitates the separation of magnitude and phase components in the signal. Adapting the Fourier magnitude profile, which characterizes the strength of spectro-temporal modulations present in the signal, separately from its phase profile, which characterizes the relative timing of these modulations, is akin to processing an *abstracted* representation of the signal, an important component of object-based attention [2], [4], [30], [31].

Based on this knowledge, task-driven adaptation strategies that improve the separation between foreground speech and background nonspeech sounds are particularly attractive for the challenge of speech activity detection (SAD). SAD refers to the task of assigning a speech or nonspeech label to samples in an observed audio recording and is a fundamental first step in a number of automated sound processing applications. For example, in automatic speech recognition tasks, one should transcribe only speech events in the observed audio. In low-noise environments with a close-talking microphone, SAD can usually be solved using traditional measures like signal energy, zero-crossing rate, and pitch (see, e.g., [32]–[34]), but performance rapidly degrades in noisy, reverberant, and free-field microphone conditions. However, research in the past decade has begun to focus on the issue of noise-robust SAD, with successful approaches leveraging prior knowledge about the differences in acoustic profiles that differentiate speech and nonspeech sounds. Early efforts based on statistical models (see, e.g., [35], [36]) have been especially improved in recent studies [37]–[47]. Furthermore, a variety of approaches have been designed to specifically exploit the spectral and temporal structure of speech [48]–[56]. Most recently, data-driven approaches based on recurrent and deep neural networks have further pushed the state of the art, yielding extraordinary results on difficult corpora [57]–[62].

In this paper, we take a different approach to SAD, and consider to what extent task-driven adaptive retuning of STRFs can improve SAD performance. We begin by reviewing relevant concepts regarding auditory peripheral and central processing as they pertain to the framework presented here. Next, we describe a computational model of task-driven STRF plasticity in the modulation domain inspired by auditory neurophysiology and explore its application to the challenge of detecting speech in noisy environments. We show how to induce adaptation in an ensemble of neurophysiological STRFs, and demonstrate how the STRFs reorient themselves to enhance the spectro-temporal modulations of speech while suppressing those associated with a variety of nonspeech sounds. Importantly, we demonstrate how features derived from the adapted STRFs improve performance in a SAD task in unseen noise conditions with respect to an unadapted ensemble and a recently proposed baseline. Lastly, to better understand how STRF adaptation affects the representational quality of a target speech sound, we consider a stimulus reconstruction task similar to those recently considered in a variety of neurophysiological studies. We show that stimuli reconstructed from STRFs adapted using our proposed framework yield a higher fidelity representation for speech in clean and additive noise conditions using a variety of objective and perceptual measures. Overall, the results suggest that a framework for task-driven adaptation formulated in the modulation domain can yield a high-fidelity, noise-robust, and improved representation of the target source that is applicable to automated speech processing tasks.

II. Preliminaries

The ascending auditory pathway comprises a hierarchy of stages that transform sound observed at the outer ear to neural responses in central cortical areas. We begin by briefly describing the relevant aspects of this processing pipeline as it pertains to the adaptation framework considered in this paper.

A. Auditory Peripheral Processing

To account for the transformation of sound from the outer ear through the auditory midbrain, we use a computational model of mammalian auditory periphery to obtain a time-frequency representation for input stimuli referred to as an *auditory spectrogram* [63]. This model accounts for peripheral processing spanning the cochlea through the auditory midbrain. First, an input signal is processed by a bank of 128 constant-Q gammatone-like filters. The filters are uniformly spaced along the logarithmic tonotopic axis, starting at 90 Hz, and span 5.3 octaves. This is followed by a first-order derivative and half-wave rectification to model lateral inhibition in the cochlear nucleus and acts to sharpen the filter responses in each channel. Finally, the responses are smoothed in time using an exponentially decaying filter with a 10 ms time constant to model short-term integration and the loss of phase locking in the midbrain. Examples of auditory spectrograms for speech and a nonspeech jet sound are shown in Fig. 1.

B. Cortical Processing Via Spectro-Temporal Receptive Fields

Beyond the midbrain, the time-varying tonotopic signal is further analyzed by ensembles of neurons in primary auditory cortex (A1) [64]. Cortical neurons essentially act as filters that extract information about the frequency content and spectro-temporal dynamics of an input auditory spectrogram, and each filter's tuning characteristics is described by its spectro-temporal receptive field [8]. As illustrated in Fig. 1, STRFs reflect sensitivity to a variety of input energy patterns, with simple shapes preferring highly localized and narrowband input to complex shapes preferring spectral, temporal, and joint spectro-temporal variations.

In this paper, we consider ensembles of *neurophysiological* STRFs estimated from recordings from non-behaving ferret primary auditory cortex, collected in the context of studies not specifically related to the current work [15], [16], [65]. We use neurophysiological STRFs because of their inherent ability to form a rich, redundant, and over-complete neural representation that captures the span of spectro-temporal modulations that characterize natural sounds [66], [67]. All STRFs were derived from neural responses to modulated noise stimuli known as temporally-orthogonal ripple combinations (TORCs) [68]. TORCs represent a spectro-temporally rich stimuli for driving cortical neuron responses, and facilitate a mathematically tractable method for estimating the transfer function of a neuron, i.e., its STRF [13].

The STRFs spanned 5 octaves in frequency over 15 channels with starting frequencies of either 125, 250, or 500 Hz. The choice of frequency range of each STRF was determined by experimental considerations, as discussed in neurophysiological studies for which data was collected [15], [16], [65]. Furthermore, the STRFs spanned 250 ms in time over 13 bins. The

ensemble STRFs included in the current study were selected from a larger sample of neurophysiological data based on two criteria: *(i)* only STRFs with a signal-to-noise ratio (SNR) > 2.4 were included; SNR was estimated based on the variance of neural responses to different repetitions of the same stimuli using a bootstrap procedure (further details can be found in [68]). *(ii)* We sorted STRFs according to a separability index $SPI \in [0, 1]$, defined as $SPI := 1 - \sigma_1^2 / \sum_j \sigma_j^2$, where σ_i is the ith singular value for a given STRF [12], [69]. All STRFs with $SPI \geq 0.5$ were removed from any further analysis. These two criteria yielded 810 neurophysiological STRFs used for the SAD experiments described later.

III. An Adaptive Framework for Speech Activity Detection

In this section, we first describe a computational framework inspired by auditory processing to induce adaptive driven changes in a set of neurophysiological STRFs. We then demonstrate how the adapted STRFs yield features that improve SAD performance across a variety of unseen noise conditions.

A. Methods

1) STRF Adaptation Framework—In earlier work, we explored a mathematical foundation for task-driven changes in neurophysiological STRFs [31]. Here, we build on these concepts to develop a model of task-driven STRF adaptation to reliably detect speech in noisy environments. The framework considered in this study is designed to be consistent with neural circuits thought to induce adaptive changes in cortical STRFs [70], and a schematic of the process is shown in Fig. 1. In essence, we model adaptation as an *iterative* process that alternates between (1) STRF perturbations that improve discrimination between speech and nonspeech sounds and (2) updates to the parameters of a linear discriminative model.

We first model the influence of top-down adaptation as the assignment of a behaviorally relevant categorical label $y_m \in \{+1, -1\}$ to an observed ensemble response $\mathbf{R}_m$, where $y_m = +1$ is associated with examples of speech and $y_m = -1$ is associated with examples of nonspeech. To improve discrimination between speech and nonspeech, we assume that the adaptive feedback acts to maximize the conditional likelihood of the labels y_m defined as

$$p(Y_m = y_m \mid \mathbf{R}_m, \mathbf{w}) := \sigma(y_m \mathbf{w}^T \mathbf{R}_m)$$

where $\sigma(\alpha) = 1/(1 + \exp(-\alpha))^{-1}$ is the logistic function, $\mathbf{w} = [w_0, w_1, \cdots, w_K] \in \mathbb{R}^{K+1}$ is a vector of regression coefficients, and K denotes the size of the neural ensemble. To obtain an ensemble response $\mathbf{R}_m$, we first define the firing rate of an individual neuron as

$$r_k(t, f; m) = h_k^A(t, f) *_{tf} s_m(t, f) \quad (1)$$

with corresponding modulation domain representation

$$|R_k(\omega,\Omega;m)| = |H_k^A(\omega,\Omega)| \cdot |S_m(\omega,\Omega)|$$

where $*_{tf}$ denotes 2D convolution in time and frequency, and $R_k(\omega, \Omega; m)$, $H_k^A(\omega, \Omega)$, and $S_m(\omega, \Omega)$ are the 2D Discrete Fourier Transforms (DFT) of firing rate, STRF, and the m'th stimulus token, respectively. ω characterizes modulations along the temporal axis (*rate*, in Hz) whereas Ω characterizes modulations along the spectral axis (*scale*, in cycles/octave). Finally, the ensemble response vector is constructed as $\mathbf{R}_m = [1, \Sigma_{\omega\Omega}|R_1(\omega, \Omega)|, \cdots, \Sigma_{\omega\Omega}|R_K(\omega, \Omega)|] \in \mathbb{R}^{K+1}$.

To induce adaptation in an ensemble of STRFs, we define the cost function

$$J(\mathbf{w}, \mathscr{H}_A) := \underbrace{\frac{1}{2}\|\mathbf{w}\|_2^2 - C \cdot \langle \log \sigma(y_m \mathbf{w}^T \mathbf{R}_m) \rangle_m}_{\text{Discriminability}} + \underbrace{\frac{\lambda}{2} \sum_k \|\Delta_k\|_F^2}_{\text{Stability}} \quad (2)$$

where $\mathscr{H}_A := \{|H_k^A(\omega,\Omega)|\}_{k=1}^K$ and $\Delta_k := |H_k^0(\omega,\Omega)| - |H_k^A(\omega,\Omega)|$. Thus, the overall goal is to determine settings of $\mathbf{w}$ and $\mathscr{H}_A$ that optimize the proposed cost function. The discriminability term is a common form of logistic regression regularized by a Gaussian prior on the weight vector $\mathbf{w}$ [71], and quantifies the average (negative) conditional log-likelihood of the labels. We also include a stability term to ensure that the adapted STRFs do not vary "too far" from their nominal tuning shape [72]. The hyperparameters C and λ allow the user to vary the influence of each term in the optimization, i.e., increasing C in the objective function will favor improved discrimination, whereas increasing λ will resist changes to the STRFs.

One strategy for jointly optimizing the shapes of the STRF modulation profiles $\mathscr{H}_A$ and regression parameters $\mathbf{w}$ is the use of block coordinate descent where we optimize Eq. (2) by alternating between solving two convex subproblems:

$$\arg \min_{\mathbf{w}} \quad J(\mathbf{w}, \mathscr{H}_A) \text{ s.t. } w_k > 0 \quad \text{(P1)}$$

$$\arg \min_{\mathscr{H}_A} \quad J(\mathbf{w}, \mathscr{H}_A) \text{ s.t. } |H_k^A(\omega,\Omega)| \geq 0, \quad \forall k, \omega, \Omega \quad \text{(P2)}$$

for $k = 1, 2, \cdots, K$. The constraints on (P1) are justified below whereas the constraints on (P2) are required since modulation profiles $|H_k^A(\omega, \Omega)|$ are necessarily nonnegative. Because $J(\mathbf{w}, \mathscr{H}_A)$ is a sum of convex functions, and the constraints on (P1) and (P2) are convex, each subproblem is therefore convex with a unique global minimum. Furthermore, since each

update to $\mathbf{w}$ and $\mathcal{H}_A$ does not increase the value of $J(\mathbf{w}, \mathcal{H}_A)$, alternating updates to $\mathbf{w}$ and $\mathcal{H}_A$ guarantee convergence to a local minimum of the overall objective function [73], [74].

The solutions to (P1) and (P2) are found numerically [75], [76] by searching for stationary points of the respective objective functions, i.e., when $\nabla_{\mathbf{w}} J(\mathbf{w}, \mathcal{H}_A) = 0$ and $\nabla_{h_k^A(t,f)} J(\mathbf{w}, \mathcal{H}_A) = 0$. For the regression coefficients, upon convergence of (P1), and assuming the minimum lies within the feasible set formed by the constraints on the w_k, the regression coefficient vector can be written as

$$\mathbf{w} = C\langle y_t \cdot [1 - \sigma(y_t \mathbf{w}^T \mathbf{r}_t)] \cdot \mathbf{r}_t \rangle_t$$

We interpret the term $[1 - \sigma(y_t \mathbf{w}^T \mathbf{r}_t)]$ as a "prediction error" and consequently hard-to-predict responses have more influence on choice of the optimal regression coefficients. Moreover, because the w_k for $k > 0$ are constrained to be positive, those coefficients can be thought of as a *population gain vector* that applies more weight to task-relevant vs. task-irrelevant neurons.

Next, upon convergence of (P2), and assuming the minimum lies within the feasible set formed by the constraints on $| H_k^A(\omega, \Omega) |$, the adapted STRF modulation profiles can be written as

$$| H_k^A(\omega, \Omega) | \; = \; | H_k^0(\omega, \Omega) | \; + \frac{C}{\lambda} \cdot w_k \cdot \langle y_m \cdot [1 - \sigma(y_m \mathbf{w}^T \mathbf{R}_m)] \cdot \; | S_m(\omega, \Omega) | \rangle_m \quad (3)$$

Eq. (3) shows how the STRF adaptation patterns are consistent with a contrast-matched filter in the modulation domain. First, task-driven STRF plasticity directly reflects the spectro-temporal modulation profiles of the speech and nonspeech stimuli, as shown in the averaging term. The impact of each stimulus sample on adaptation is proportional to the difficulty of predicting its corresponding label. Importantly, because we have constrained the regression coefficients w_k to be positive, we are guaranteed that speech modulations are *enhanced* whereas those from nonspeech are *suppressed*. Finally, the first term acts to resist changes from the initial STRF modulation profile, with the magnitude of the effect being controlled C by and λ.

Upon optimizing the cost function in Eq. (2), we obtain a set of adapted modulation profiles $\mathcal{H}_A$. To analyze the effect of STRF adaptation on the neural ensemble, we consider the average difference between the adapted and initial modulation profiles by computing

$$\Delta MTF(\omega, \Omega)\rangle := \langle \; | H_k^A(\omega, \Omega) | \; - \; | H_k^0(\omega, \Omega) | \; \rangle_k$$

In this way, one can visualize which modulations of speech and nonspeech stimuli are enhanced or suppressed, respectively.

2) Proposed SAD System—To test the hypothesis that task-adapted STRFs can improve speech activity detection in unseen noisy environments, we consider a Gaussian Mixture Model (GMM)-based SAD system and compare performance of between features derived from the initial and adapted STRFs. An overview of the proposed SAD system is shown in Fig. 2. For training (top row), we use clean speech and a variety of nonspeech samples to extract a set of features from the passive ensemble $\mathcal{H}_0$, yielding ensemble responses $\{r_k^0(t,f)\}_k$. Feature extraction is followed by a series of post-processing and dimensionality reduction steps. First, the responses are full-wave rectified and averaged over one-second intervals every 50 ms, yielding three-dimensional tokens $\mathbf{R}_0(t', f, k)$. Next, we apply dimensionality reduction using the tensor singular value decomposition (TSVD) [77], projecting the tokens to a subspace that retains 99.9% of the variance along each dimension, and stack the reduced-dimension tokens to obtain column vectors $\mathbf{r}_{t'}^0$. We then standardize each vector to have zero-mean and unit-variance. Finally, we finally fit GMMs to the observed speech and nonspeech tokens, yielding model parameters (i.e., weights, means, and covariance matrices) Θ_S and Θ_{NS}, respectively.

For testing (Fig. 2, bottom row), features are similarly extracted from observed noisy speech utterances using both passive and adapted STRFs $\mathcal{H}_0$ and $\mathcal{H}_A$, respectively. We again apply post-processing and dimensionality reduction, and compute the log-likelihood ratio (LLR) of speech versus nonspeech using the GMMs trained in the passive conditions. We evaluate system performance by sweeping a threshold on the LLR scores, labeling tokens that exceed the threshold as speech, and those below as nonspeech. Using ground truth labels, for a given threshold value we compute miss and false alarm probabilities, p_M and p_{FA}, respectively, and consider these error probabilities across all thresholds to yield a detection error tradeoff (DET) curve [78]; an example DET curve is shown ahead in Fig. 4(A). To summarize performance of the system, we compute the equal-error rate (EER), i.e., the threshold setting that yields $p_M = p_{FA}$. A system that performs well has small p_M and p_{FA} across a broad range of thresholds, hence (1) the corresponding DET curve will be close to the origin and (2) EER will be small.

B. Experimental Setup

1) STRF Adaptation Framework—We use clean speech from the TIMIT corpus [79] and `white`, `babble`, `street`, `restaurant`, and `f16` noise from the NOISEX-92 corpus [80]. The speech class is constructed to contain equal amounts of clean and noisy speech (at 5 dB SNR) whereas the nonspeech class contains equal amounts of pure noise from the five noise classes. The use of clean and noisy speech reflects the notion that a listener has prior knowledge of speech in both clean and moderately noisy environments. We use audio samples approximately 3 seconds in length, apply pre-emphasis, and standardize each waveform to be zero-mean and unit variance, and we use approximately 5 minutes of audio for each class. Next, for each audio sample we compute an auditory spectrogram, apply cube-root compression, and downsample along the frequency axis to 32 channels. Finally,

the 2D discrete Fourier Transform (DFT) was applied to 250 ms segments of spectrogram, followed by the modulus operation to obtain input tokens $|S_m(\omega, \Omega)|$ for the adaptation algorithm. Tokens are scaled to have unit variance for each class as this seemed to improve convergence time of the adaptation algorithm.

We next select random samplings of $K = 50$ STRFs from the large physiological ensemble described in Section II-B. The use of a larger ensemble is computationally challenging because the number of parameters involved in the optimization of (P2) becomes prohibitively large. Furthermore, we find that random samplings 50 STRFs are sufficient to tile the relevant modulation space of speech tokens given the redundant and overcomplete nature of neurophysiological receptive fields [66], [67]. Next, each STRF is interpolated along the time and frequency axis to match the temporal and spectral sampling of the input tokens (i.e., 100 Hz temporal and 6.04 cyc/oct spectral, respectively). We also assume that each STRF has a starting frequency of 90 Hz spanning 5.3 octaves. We scale each STRF to have unit Euclidean norm, and apply the 2D DFT, followed by the modulus operation, to obtain the initial set of modulation profiles $\mathscr{H}_0 := \{ | H_k^0(\omega, \Omega) | \}_{k=1}^{K}$. This set represents a "passive" listening state, one where adaptation is not induced. Note that for adaptation we only need to consider the first two quadrants of the DFT since for real-valued input $| H_k^0(\omega, \Omega) | = | H_k^0(-\omega, -\Omega) |$. Finally, to visualize the adapted STRFs in the original time-frequency domain, we use the phase of the original passive filters.

2) Proposed SAD System—Firing rates are computed as in Eq. (1), with 128-channel auditory spectrograms and cube-root compression applied. The STRFs $h_k(t, f)$ are also interpolated to span the full 128 channels. Post-processing and dimensionality reduction are performed as described in Section III-A.

For the passive and adapted STRFs, we consider three random draws from the large physiological STRF set as well as a range[1] of model hyperparameters C; we report results for the STRF ensemble that yields the best performance. We train our GMM SAD system using clean speech from the TIMIT corpus and nonspeech samples from the BBC Sound Effects Library [81]. The nonspeech set comprises an equal amount of audio from a range of acoustic classes[2]. For both the speech and nonspeech categories, we use 7500 one-second tokens, or approximately 2.1 hrs of audio, and reduce the dimension of extracted features via TSVD from (128 frequency channels ×50 STRFs) to 40-dimensional column vectors. Finally, we fit 32-mixture GMMs to the speech and nonspeech categories.

We evaluate our system in unseen noise cases using audio from the QUT-NOISE-TIMIT corpus, a database specifically designed for evaluating different SAD algorithms [82]. The corpus is formed by mixing clean TIMIT utterances with realistic background noise covering four noise scenarios[3] (`STREET`, `CAFE`, `CAR`, and `HOME`) with two unique locations

[1]We found it was sufficient to fix λ and explore a range of values for C.
[2]For this study we used the `Emergency`, `Foley`, `Industry and Machines`, `Technology`, `Transportation`, and `Water` classes.
[3]There is also a fifth `REVERB` condition, but for our experiments it is ignored.

per noise type at various SNRs. For our experiments, we select ten utterances (each 60-seconds in length) at random from each noise condition and SNR, ensuring that there is no overlap between TIMIT utterances seen in training and those used in testing. For a baseline comparison, we use the statistical model-based likelihood ratio test of Tan *et al.* that leverages harmonicity cues to improve detection of speech in noisy environments [40]; this approach has been shown to work well in a variety of noise conditions.

C. Results

1) STRF Adaptation Framework—We first apply the model to simulate a scenario where a listener adapts processing to focus on speech sounds in additive noise environments. This result is shown in Fig. 3(A), and illustrates that the overall effect of task-driven adaptation is to increase population sensitivity to slower modulations, with the effect being stronger for downward vs. upward moving modulations (i.e., the right- vs. the left-half planes, respectively), while suppressing sensitivity to faster modulations away from the origin. This pattern was also found for other random selections of the initial ensemble $\mathcal{H}_0$. The magnitude of the effect depends on choice of model hyperparameters, becoming stronger for decreasing λ and increasing C (data not shown). Finally, shown next in panels B and C are the adaptation patterns of two individual neurons, illustrating how the neurons broaden and reorient themselves to better focus on upward and downward modulations as suggested by panel A.

2) Proposed SAD System—Shown in Fig. 4(A) is an example DET curve for the `STREET` noise for the baseline and proposed system using passive and adapted STRFs. The DET curve is obtained by pooling LLR scores across SNRs dB. $\in \{-5, +5, +10, +15\}$dB. Because the DET curve for the adapted ensemble $\mathcal{H}_A$ is closest to the origin with no overlap with the other curves, it represents clear improvement over the baseline and the passive ensemble $\mathcal{H}_0$ across all SNRs, with absolute reductions in EER of approx. 11% and 3%, respectively.

To better understand how the adapted STRFs improve performance, Panel B shows an analysis of the distribution of LLR scores for the STRF ensembles with respect to speech and nonspeech categories. These results show that under the GMMs trained on the passive STRF features, use of the adapted STRF ensemble increases the overall likelihoods of speech and nonspeech. However, despite this added bias to the scores, there is an overall improved separation between the speech and nonspeech distributions as computed using the Kullback-Leibler divergence (KL_0 =59.8 KL_A = 76.2 and assuming Gaussian-distributed scores). Similar improvements are found across the other noise scenarios.

Finally, panel C summarizes the overall performance of the various SAD configurations in terms of average EER across all noise conditions, showing that the adapted STRFs improve over the baseline and passive STRF results by an absolute 8% and 3.5%, respectively.

IV. Stimulus Reconstruction

The previous section showed empirically that the proposed STRF adaptation framework improved detection of speech in unseen noisy conditions by increasing the separability

between the LLR scores of speech and nonspeech (Fig. 4(B)). However, we sought to better understand how the adapted STRF ensemble improved the representational fidelity of a target speech signal. One way to assess the ability of a neural ensemble to encode features of the target source is to use a stimulus reconstruction approach. By reconstructing the observed input to the ensemble, one can assess how the features of the input are encoded by the population. One can then vary the state of the ensemble (i.e., passive vs. adapted) and compare the reconstructions with the original stimulus.

The stimulus reconstruction approach has been successful in a number of neurophysiological studies. The approach was pioneered in studies of the fly visual system [83], [84] and has been used to study feature encoding in visual [85], [86] and auditory cortical circuits [87], [88]. Of particular interest are recent studies that have shed light on how cortex represents imagined speech [89] and the nature of how top-down adaptive feedback influences representation in cortical circuits [90]–[92].

In this section, we explore the stimulus reconstruction approach as it relates to the challenge of speech-in-noise detection, using the approach outlined by Mesgarani *et al.* [87]. We hypothesize that adapting an STRF ensemble according to the proposed model yields a higher fidelity representation of the target stimulus, and we explore this via reconstruction experiments in clean and additive noise conditions using objective and perceptual measures.

A. Stimulus Reconstruction Model

In physiological studies, neural firing rate is typically modeled as

$$\widetilde{r}_k(t) = \sum_{f=1}^{F} h_k(t,f) *_t s(t,f)$$

where $h_k(t, f)$ is an STRF, $s(t, f)$is the stimulus, F is the number of frequency channels, $t \in [1, T]$, and $*_t$ denotes convolution in time. To reconstruct an input stimulus from observed neural firing rates we use the linear form

$$\hat{s}(t,f) = \sum_{k=1}^{K} \sum_{\tau=1}^{\tau_M} g(k,\tau,f)\widetilde{r}_k(t-\tau) \quad (4)$$

where $\tau_M > 0$ is the (user-defined) temporal extent of the reconstruction filters $\{g(k, \tau, f)\}$ and is a collection of inverse mapping functions [87].

Eq. (4) implies that individual frequency channels are independent of one another and hence we can compactly write

$$\hat{s}_f(t) = \sum_{k=1}^{K} \sum_{\tau=1}^{\tau_M} g_f(\tau; k) \tilde{r}_k(t-\tau)$$

$$= \mathbf{g}_f^T R$$

where

$$\mathbf{g}_f = \mathrm{vec}\left[\begin{pmatrix} g_f(1,1) & g_f(1,2) & \cdots & g_f(1,K) \\ g_f(2,1) & g_f(2,2) & \cdots & g_f(2,K) \\ \vdots & \vdots & & \vdots \\ g_f(\tau_M,1) & g_f(\tau_m,1) & \cdots & g_f(\tau_M,K) \end{pmatrix}\right]$$

and

$$R = \begin{pmatrix} \tilde{r}_1(1) & \tilde{r}_1(2) & \cdots & \tilde{r}_1(\tau_M) & \cdots & \tilde{r}_1(T) \\ 0 & \tilde{r}_1(1) & \cdots & \tilde{r}_1(\tau_M - 1) & \cdots & \tilde{r}_1(T-1) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \tilde{r}_t(1) & \cdots & \tilde{r}_1(T-\tau_M) \\ \tilde{r}_2(1) & \tilde{r}_2(2) & \cdots & \tilde{r}_2(\tau_M) & \cdots & \tilde{r}_2(T) \\ 0 & \tilde{r}_2(1) & \cdots & \tilde{r}_2(\tau_M - 1) & \cdots & \tilde{r}_2(T-1) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \tilde{r}_2(1) & \cdots & \tilde{r}_2(T-\tau_M) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \tilde{r}_K(1) & \cdots & \tilde{r}_K(T-\tau_M) \end{pmatrix}$$

The vec(·) operator performs a column-wise stacking of the input matrix. Furthermore, defining the inverse mapping matrix $G := [\mathbf{g}_1 \cdots \mathbf{g}_F]$, we can write $\hat{S} := \hat{s}(t, f) = G^T R$

One way to arrive at an optimal reconstruction of $\hat{S}$ is to determine the matrix G^* that solves the least-squares problem

$$G^* = \arg\min_G \|S - \hat{S}\|_F^2$$

where $S := s(t, f)$ is the observed stimulus. This closed-form solution is readily obtained as

$$G^* = C_{RR}^{-1} C_{RS} \quad (5)$$

where $C_{RR} = RR^T$ is the response autocorrelation matrix and $C_{RS} = RS^T$ is the stimulus-response correlation matrix. Because some of the STRFs in the neural ensemble have similar shapes, the observed firing rates consequently yield redundancies in the rows of R. Thus, it is often the case that C_{RR} is poorly conditioned, necessitating the use of some form of regularization to properly invert the response autocorrelation matrix. Here we use the subspace regression approach proposed by Theunissen *et al.* [11]. Since C_{RR} is real-symmetric, it can be expressed as $C_{RR} = U\Sigma U^T$ where U is a matrix of eigenvectors, $\Sigma = \text{diag}(\sigma_1, \cdots \sigma_{K \cdot \tau_M})$, and σ_i are the eigenvalues of C_{RR}. Because of the redundancies in the rows of R, it generally holds that rank $(C_{RR}) < K \cdot \tau_M$, and thus we should ignore the eigenvectors corresponding to small eigenvalues (otherwise they tend to introduce noise once C_{RR} is inverted). We set eigenvalues smaller than a pre-defined threshold η to zero.

B. Experimental Setup

We consider two measures to evaluate the quality of a given reconstruction from an inverse mapping obtained by Eq. (5). The first is the temporally averaged mean-square error between the original and reconstructed spectrogram, defined as $MSE := \langle \|s(t,f) - \hat{s}(t,f)\|_F^2 \rangle_t$ and serves as an objective measure of reconstruction quality. The second is a perceptual comparison between the original time-domain waveform and synthesized version obtained using the Gaussian convex projection algorithm [26], [63]. The comparison between the waveforms is made using the ITU standard Perceptual Evaluation of Speech Quality (PESQ) measure [93]. PESQ ranges between 1 and 5 and correlates well with listener-reported mean opinion scores of perceptual quality, with higher scores indicating higher quality.

To study how reconstruction performance varies as a function of adaptation state, we use the passive and adapted STRF ensembles $\mathcal{H}_0$ and $\mathcal{H}_A$ to obtain optimal inverse mappings G_0 and G_A, respectively. We use 350 clean speech utterances from the TIMIT `train` corpus (approx. 17.5 minutes) to learn the inverse mapping matrices. The neural responses $\tilde{r}_k(t)$ are also standardized to have zero-mean and unit variance prior to obtaining and G_0 and G_A, and these parameters are applied to subsequent reconstructions. For a given inverse mapping, we also consider a range of inverse filter lengths spanning $\tau_M \in [50, 750]$ ms and eigenvalue thresholds $\log_{10} \eta \in \{-9, -6, -3\}$. Results are reported here for ensembles that achieve minimum average mean-square reconstruction error on a test set of 100 clean speech utterances from the TIMIT `test` corpus. For synthesizing time-domain waveforms, we first apply a max $(\cdot, 0)$ nonlinearity to synthesized spectrograms, followed by a maximum of 30 iterations of the Gaussian convex projection algorithm.

C. Reconstruction Results

Shown in Fig. 5(A) are examples of reconstructions from clean utterances obtained using the passive and adapted STRF ensembles. We first find that both reconstructions are somewhat noisy, with G_0 yielding a distinct temporal distortion whereas G_A introduces spurious patches of spectro-temporal energy. However, both reconstructions are sufficient to capture the broad prosodic characteristics of the reference spectrogram, with good qualitative matches between pitch variations, syllabic rate, and broad formant dynamics over time. Furthermore, it is clear that G_A yields a reconstruction with better spectral resolution, since

the harmonic peaks during sections of voicing are far more pronounced as compared with G_0. Next, Panel B shows that across the test set, the adapted ensemble yields an objectively better reconstruction, with G_A yielding a significantly lower reconstruction error as compared to G_0(t-test, $p \approx 0$). Finally, the perceptual analysis in Panel C shows G_A yields a significantly higher quality waveform synthesis compared G_0 (t-test, $p \approx 0$). Of course, while PESQ values between 2–3 are generally considered somewhat noisy, informal listening tests confirm that the synthesized waveforms are nevertheless intelligible, G_A with conveying a better percept of voicing and pitch.

We also explore the extent to which the passive and adapted STRF ensembles encode information about the attended source in additive noise conditions. Here we consider a test set of 100 utterances from the TIMIT `test` corpus corrupted by additive noise from the NOISEX-92 corpus at a variety of SNRs. In addition to the `babble` and `street` noises used in training, we also consider the unseen noise classes `airport, buccaneer1` (a type of fighter jet), `factory1`, and `m109` (a type of tank). For each noisy utterance we reconstruct the spectrogram using the inverse mappings and G_0 and G_A. We then quantify reconstruction quality using the time-averaged mean-squared error between the clean reference spectrogram and the noisy reconstruction. These results, averaged across all noise types, are shown in Fig. 6. It is clear that while reconstruction quality degrades with increasing noise, in all SNR cases the adapted ensembles yield, on average, a higher quality reconstruction with respect to the clean references. We find no differences between average MSE for the seen vs. unseen noise cases.

In summary, the results of this section suggest that the proposed model of task-driven adaptation induces STRF changes that a facilitate higher-fidelity representation of attended speech in clean and noisy environments. This lends further insight as to how such an adaptation strategy is able to improve detection of speech in noisy environments.

V. Discussion

In this paper, we applied a model of auditory receptive field plasticity that acts in the modulation domain to simulate a listener dynamically adapting cognitive resources to better track speech in noisy acoustic environments. We first described how an ensemble of initial STRFs adapt to highlight the differences in spectro-temporal modulation profiles for speech vs. nonspeech sounds. We showed that the model induces STRF plasticity that strengthens relatively slow spectro-temporal modulations close to the origin (in the rate-scale domain) while simultaneously suppressing faster modulations away from the origin. We then showed how use of the adapted STRFs improves the separation between the representation of speech and nonspeech sounds, resulting in a substantial performance gain in a speech activity detection task across a variety of previously unseen noise types. Finally, we explored, via stimulus reconstruction experiments, the extent to which the passive and adapted STRF ensembles captured the salient features of the target speech source. These results showed how the use of task-driven adaptation can improve the representation of a speech target in clean and noisy conditions, as confirmed by objective and perceptual measures. This helped shed light as to how the representation improved to help facilitate the detection of speech in noise. The overall results suggest that STRFs adapted according to a biologically motivated

model of task-driven adaptation can form a noise-robust representation of sound that is applicable to automated speech processing applications.

Our model predicts that targeting speech versus nonspeech distractors enhances sensitivity of STRFs to “slower” spectro-temporal modulations. This is illustrated by the average difference modulation profile in Fig. 3. Increased slowness in the modulation domain is realized in the time-frequency domain as an overall broadening and reorientation of the STRFs, and reflects an enhancement of the modulations known to characterize speech and other natural sounds [11], [25], [26], [28]. The fact that we obtain improved SAD results using “slower” filters is consistent with other strategies that concentrate the feature extraction pipeline to the range of speech-specific modulations [27], [94]–[96]. Moreover, the STRF adaptation patterns observed here are broadly compatible with traditional signal processing schemes that emphasize slow modulations for improving noise robustness in speech tasks [97]. The distinction here is that our approach adapts the filter shapes “on the fly” and, as our SAD results suggest, such changes can be compatible with an existing statistical model to improve task performance.

As stated earlier, tracking speech in noisy environments is a critical component of human communication, and is a task at which listeners are especially adept. Based on this, and the fact that reliably detecting speech-containing regions is a critical first step in many common speech applications, the task of speech activity detection is a natural fit for our framework. Our goal in this study was not to build a state-of-the-art SAD system *per se*, but to instead focus on the design and understanding of the impact of an adaptive spectro-temporal algorithm for speech that is grounded in auditory neurophysiology. Admittedly, the proposed framework is computationally expensive, particularly for the STRF modulation profile adaptation described in (P2), which involves an optimization over $O(K \cdot T \cdot F)$ free parameters (where T and F are the number of temporal and spectral bins in the STRFs, respectively) using a complex interior-point numerical solver [76]. However, we expect that performance of the framework would improve by optimizing choice of the initial filter set, or perhaps by approximating the adapted STRF profiles based on the average modulation profile difference $\langle \Delta MTF \rangle$; we leave this fine-tuning of the system for future work. That being said, many robust approaches to SAD have been proposed over the years that perform well in noisy and reverberant environments (see, e.g., [32], [35]–[37], [40]). In particular, when large amounts of training data are available, extraordinary results can be achieved for the SAD task on difficult corpora using high-order GMMs [45] and convolutional neural networks [61], [98]. Nevertheless, these systems continue to be challenged by nonstationary interference, mismatch between expected and observed acoustics, and limited training data, and it is our hope that the modulation-domain adaptation framework presented here can be leveraged to improve these approaches.

More generally, the increasing availability of mobile sound processing applications has resulted in a significant increase in the variety of acoustic environments, communication channels, and noise conditions encountered by existing systems. Consequently, this necessitates signal processing strategies that must gracefully accommodate these factors to maintain state-of-the-art performance. We contend that because nature has converged to a robust solution for handling unseen and noisy acoustics, there is much to leverage from

auditory neurophysiology when designing automated sound processing systems. Generally speaking, cortically inspired feature representations based on spectro-temporal receptive fields underlie a number of successful approaches to noise robust speech activity detection [27], speech and speaker recognition [94]–[96], [99], and auditory scene classification [100]. The present study, in concert with other recent work in our lab [31], [101], [102], represents an extension of this methodology by incorporating the cognitive effects of dynamic, task-driven sensory adaptation as part of the feature extraction pipeline. It is our belief that new and existing systems can only benefit by incorporating the adaptive mechanisms as outlined in this paper.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant IIS-0846112, in part by the National Institute of Health under Grant 1R01AG036424, and in part by the Office of Naval Research under Grants N000141010278 and N00014-12-1-0740.

Many thanks to Shihab Shamma, Jonathan Fritz, and the Neural Systems Laboratory at the University of Maryland, College Park for providing the ferret STRFs used in this work.

References

1. Fritz JB, Elhilali M, David SV, Shamma SA. Auditory attention– focusing the searchlight on sound. Current Opinion Neurobiol. 2007; 17(4):437–455.
2. Shinn-Cunningham BG. Object-based auditory and visual attention. Trends Cognitive Sci. 2008; 12(5):182–186.
3. Bajo VM, King AJ. Focusing attention on sound. Nature Neurosci. 2010; 13(8):913–914. [PubMed: 20661266]
4. Bizley JK, Cohen YE. The what, where and how of auditory-object perception. Nat Rev Neurosci. 2013; 14(10):693–707. [PubMed: 24052177]
5. Shamma S, Fritz J. Adaptive auditory computations. Current Opinion Neurobiol. 2014; 25(0):164–168.
6. Thompson, R. Neural Mechanisms of Goal-directed Behavior and Learning. Amsterdam, The Netherlands: Elsevier; 1980.
7. Aboitiz, F., Cosmelli, D. From Attention to Goal-Directed Behavior Neurodynamical, Methodological and Clinical Trends. New York, NY, USA: Springer; 2009.
8. Aertsen AMHJ, Johannesma PIM. The spectro-temporal receptive field. Biol Cybern. 1981; 42:133–143. [PubMed: 7326288]
9. Eggermont JJ, Aertsen AM, Johannesma PI. Quantitative characterisation procedure for auditory neurons based on the spectrotemporal receptive field. Hear Res. 1983; 10(2):167–190. [PubMed: 6602799]
10. Aertsen AMHJ, Johannesma PIM. Spectro-temporal receptive fields of auditory neurons in the grassfrog. I. characterization of tonal and natural stimuli. Biol Cybern. 1980; 38:223–234.
11. Theunissen FE, Sen K, Doupe AJ. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci. 2000; 20(6):2315–2331. [PubMed: 10704507]
12. Depireux DA, Simon JZ, Klein DJ, Shamma SA. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol. 2001; 85(3):1220–1234. [PubMed: 11247991]
13. Klein DJ, Simon JZ, Depireux DA, Shamma SA. Stimulus- invariant processing and spectrotemporal reverse correlation in primary auditory cortex. J Comput Neurosci. 2006; 20(2): 111–136. [PubMed: 16518572]
14. Wooley SMN, Gill PR, Fremouw T, Theunissen FE. Functional groups in the avian auditory system. J Neurosci. 2009; 20(9):2780–2793.

15. Fritz J, Shamma S, Elhilali M, Klein D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nature Neurosci. 2003; 6(11):1216–1223. [PubMed: 14583754]
16. Fritz JB, Elhilali M, Shamma SA. Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks. J Neurosci. 2005; 25(33):7623–7635. [PubMed: 16107649]
17. Fritz JB, Elhilali M, Shamma SA. Adaptive changes in cortical receptive fields induced by attention to complex sounds. J Neurophysiol. 2007; 98(4):2337–2346. [PubMed: 17699691]
18. Beitel RE, Schreiner CE, Cheung SW, Wang X, Merzenich MM. Reward-dependent plasticity in the primary auditory cortex of adult monkeys trained to discriminate temporally modulated signals. Proc Nat Acad Sci. 2003; 100(19):11 070–11 075.
19. David SV, Hayden BY, Mazer JA, Gallant JL. Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. Neuron. 2008; 59(3):509–521. [PubMed: 18701075]
20. Yin P, Fritz JB, Shamma SA. Rapid spectrotemporal plasticity in primary auditory cortex during behavior. J Neurosci. 2014; 34(12):4396–4408. [PubMed: 24647959]
21. Feldman DE, Brecht M. Map plasticity in somatosensory cortex. Science. 2005; 310(5749):810–815. [PubMed: 16272113]
22. Mandairon N, Linster C. Odor perception and olfactory bulb plasticity in adult mammals. J Neurophysiol. 2009; 101(5):2204–2209. [PubMed: 19261715]
23. Gilbert CD, Li W. Adult visual cortical plasticity. Neuron. 2012; 75(2):250–264. [PubMed: 22841310]
24. Greenberg, S., Popper, A., Ainsworth, W. Speech Processing in the Auditory System. Berlin, Germany: Springer; 2004.
25. Elhilali M, Chi T, Shamma SA. A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun. 2003; 41:331–348.
26. Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Amer. 2005; 118(2):887–906. [PubMed: 16158645]
27. Mesgarani N, Slaney M, Shamma S. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Trans Audio, Speech, Lang Process. May; 2006 14(3):920–930.
28. Elliott TM, Theunissen FE. The modulation transfer function for speech intelligibility. PLoS Comput Biol. 2009; 5(3):e1000302. [PubMed: 19266016]
29. Mesgarani N, Shamma S. Denoising in the domain of spectrotemporal modulations. EURASIP J Audio, Speech, Music Process. 2007:042357.
30. Griffiths TD, Warren JD. What is an auditory object? Nature Rev Neurosci. 2004; 5(11):887–892. [PubMed: 15496866]
31. Carlin MA, Elhilali M. Modeling attention-driven plasticity in auditory cortical receptive fields. Frontiers Comput Neurosci. 2015; 9(106)
32. Benyassine A, Shlomot E, yu Su H, Massaloux D, Lamblin C, Petit J-P. ITU-T recommendation G. 729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. IEEE Commun Mag. Sep; 1997 35(9):64–73.
33. Chengalvarayan R. Robust energy normalization using speech/non-speech discriminator for german connected digit recognition," in. Proc EUROSPEECH. 1999; 99:61–64.
34. Woo K-H, Yang T-Y, Park K-J, Lee C. Robust voice activity detection algorithm for estimating noise spectrum. IEEE Electron Lett. Jan; 2000 36(2):180–181.
35. Ephraim Y, Malah D. Speech enhancement using a minimum- mean square error short-time spectral amplitude estimator. IEEE Trans Acoust, Speech, Signal Process. Dec; 1984 ASSP-32(6): 1109–1121.
36. Sohn J, Kim NS, Sung W. A statistical model-based voice activity detection. IEEE Signal Process Lett. Jan; 1999 6(1):1–3.
37. Ramirez J, Segura J, Benitez C, Garcia L, Rubio A. Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Process Lett. Oct; 2005 12(10):689–692.
38. Borgstrom B, Alwan A. Improved speech presence probabilities using HMM-based inference, with applications to speech enhancement and ASR. IEEE J Sel Topics Signal Process. Oct; 2010 4(5): 808–815.

39. Park C, Kim N, Cho J, Kim J. Integration of sporadic noise model in POMDP-based voice activity detection. Proc ICASSP. 2010:4486–4489.
40. Tan LN, Borgstrom BJ, Alwan A. Voice activity detection using harmonic frequency components in likelihood ratio test. Proc ICASSP. 2010:4466–4469.
41. Yu T, Hansen J. Discriminative training for multiple observation likelihood ratio based voice activity detection. IEEE Signal Process Lett. Nov; 2010 17(11):897–900.
42. Deng S, Han J, Zheng T, Zheng G. A modified MAP criterion based on hidden Markov model for voice activity detecion. Proc ICASSP. 2011:5220–5223.
43. Mousazadeh S, Cohen I. AR-GARCH in presence of noise: Parameter estimation and its application to voice activity detection. IEEE Trans Audio, Speech, Lang Process. May; 2011 19(4):916–926.
44. Petsatodis T, Boukis C, Talantzis F, Tan Z-H, Prasad R. Convex combination of multiple statistical models with application to VAD. IEEE Trans Audio, Speech, Lang Process. Nov; 2011 19(8): 2314–2327.
45. Ng T, Zhang B, Nguyen L, Matsoukas S, Zhou X, Mesgarani N, Veselỳ K, Matejka P. Developing a speech activity detection system for the DARPA RATS program. Interspeech. 2012
46. Suh Y, Kim H. Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection. IEEE Signal Process Lett. Aug; 2012 19(8):507–510.
47. Amehraye A, Fillatre L, Evans N. Voice activity detection based on a statistical semiparametric test. Proc ICASSP. 2013:6367–6371.
48. Bach J, Kollmeier B, Anemüller J. Modulation-based detection of speech in real background noise: Generalization to novel background classes. Proc ICASSP. 2010:41–44.
49. Dhananjaya N, Yegnanarayana B. Voiced/nonvoiced detection based on robustness of voiced epochs. IEEE Signal Process Lett. Mar; 2010 17(3):273–276.
50. Fukuda T, Ichikawa O, Nishimura M. Long-term spectro-temporal and static harmonic features for voice activity detection. IEEE J Sel Topics Signal Process. Oct; 2010 4(5):834–844.
51. Ghosh P, Tsiartas A, Narayanan S. Robust voice activity detection using long-term signal variability. IEEE Trans Audio, Speech, Lang Process. Mar; 2011 19(3):600–613.
52. McCowan I, Dean D, McLaren M, Vogt R, Sridharan S. The delta-phase spectrum with application to voice activity detection and speaker recognition. IEEE Trans Audio, Speech, Lang Process. Sep; 2011 19(7):2026–2038.
53. You D, Han J, Zheng G, Zheng T. Sparse power spectrum based robust voice activity detector. Proc ICASSP. 2012:289–292.
54. Hsu C-C, Lin T-E, Chen J-H, Chi T-S. Voice activity detection based on frequency modulation of harmonics. Proc ICASSP. 2013:6679–6683.
55. Sadjadi S, Hansen J. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process Lett. Mar; 2013 20(3):197–200.
56. Omar M, Ganapathy S. Shift-invariant features for speech activity detection in adverse radio-frequency channel conditions. Proc ICASSP. 2014:6309–6313.
57. Eyben F, Weninger F, Squartini S, Schuller B. Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies. Proc ICASSP. 2013:483–487.
58. Hughes T, Mierle K. Recurrent neural networks for voice activity detection. Proc ICASSP. 2013:7378–7382.
59. Zhang XL, Wu J. Deep belief networks based voice activity detection. IEEE Trans Audio, Speech, Lang Process. Apr; 2013 21(4):697–710.
60. Zhang X-L, Wu J. Denoising deep neural networks based voice activity detection. Proc ICASSP. 2013:653–657.
61. Thomas S, Ganapathy S, Saon G, Soltau H. Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions. Proc ICASSP. 2014:853–857.
62. Zhang X-L. Unsupervised domain adaptation for deep neural network based voice activity detection. Proc ICASSP. 2014:6864–6868.
63. Yang X, Wang K, Shamma SA. Auditory representations of acoustic signals. IEEE Trans Inf Theory. Mar; 1992 38(2):824–839.

64. Pickles, JO. An Introduction to the Physiology of Hearing. 3. Bradford, U.K: Emerald Group Publishing; 2008.
65. Elhilali M, Fritz JB, Klein DJ, Simon JZ, Shamma SA. Dynamics of precise spike timing in primary auditory cortex. J Neurosci. 2004; 24(5):1159–1172. [PubMed: 14762134]
66. Miller LM, Escabi MA, Read HL, Schreiner CE. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol. 2002; 87(1):516–527. [PubMed: 11784767]
67. Escabi MA, Read HL. Representation of spectrotemporal sound information in the ascending auditory pathway. Biol Cybern. 2003; 89(5):350–362. [PubMed: 14669015]
68. Klein DJ, Depireux DA, Simon JZ, Shamma SA. Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. J Comput Neurosci. 2000; 9(1):85–111. [PubMed: 10946994]
69. Carlin MA, Elhilali M. Sustained firing of central auditory neurons yields a discriminative spectro-temporal representation for natural sounds. PLoS Comput Biol. 2013; 9(3):e1002982. [PubMed: 23555217]
70. Shamma, S., Fritz, J., David, S., Winkowski, D., Yin, P., Elhilali, M. The Neurophysiological Bases of Auditory Perception. Vol. ch 51. New York, NY, USA: Springer; 2010. Correlates of auditory attention and task performance in primary auditory and prefrontal cortex, ser; p. 555-570.
71. Bishop, CM. Pattern Recognition and Machine Learning. New York, NY, USA: Springer; 2006.
72. Elhilali M, Fritz JB, Chi TS, Shamma SA. Auditory cortical receptive fields: Stable entities with plastic abilities. J Neurosci. 2007; 27(39):10 372–10 382.
73. Bertsekas, D. Nonlinear Programming. 2. Nashua, NH, USA: Athena Scientific; 1999.
74. Boyd, S., Vandenberghe, L. Convex Optimization. Cambridge, U.K: Cambridge Univ. Press; 2004.
75. Grant, M., Boyd, S. Graph implementations for nonsmooth convex programs. Recent Advances in Learning and Control, ser. In: Blondel, V.Boyd, S., Kimura, H., editors. Lecture Notes in Control and Information Sciences. Berlin, Germany: Springer-Verlag; 2008. p. 95-110.
76. Grant, M., Boyd, S. CVX: Matlab software for disciplined convex programming, version 2.1. Mar, 2014. [Online]. Available: http://cvxr.com/cvx
77. De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. SIAM J Matrix Anal Applicat. 2000; 21(4):1253–1278.
78. Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M. The DET curve in assessment of detection task performance. Proc EUROSPEECH. 1997:1895–1898.
79. Garofolo, JS., Lamel, LF., Fisher, WM., Fiscus, JG., Pallett, DS., Dahlgren, NL., Zue, V. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Philadelphia, PA, USA: Linguistic Data Consortium; 1993.
80. Varga, A., Steeneken, H., Tomlinson, M., Jones, D. Tech Rep. Speech Research Unit, Defense Research Agency; Malvern, U.K: 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition.
81. The BBC Sound Effects Library Original Series. 2006. [Online]. Available: http://www.soundideas.com
82. Dean DB, Sridharan S, Vogt RJ, Mason MW. The QUT-NOISE- TIMIT corpus for the evaluation of voice activity detection algorithms. Proc Interspeech. 2010
83. Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D. Reading a neural code. Science. Jun; 1991 252(5014):1854–1857. [PubMed: 2063199]
84. de Ruyter van Steveninck RR, Lewen GD, Strong SP, Koberle R, Bialek W. Reproducibility and variability in neural spike trains. Science. Mar; 1997 275(5307):1805–1808. [PubMed: 9065407]
85. Buracas GT, Zador AM, DeWeese MR, Albright TD. Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. Neuron. May; 1998 20(5):959–969. [PubMed: 9620700]
86. Miyawaki Y, Uchida H, Yamashita O, Sato M-a, Morito Y, Tanabe HC, Sadato N, Kamitani Y. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron. Dec; 2008 60(5):915–929. [PubMed: 19081384]
87. Mesgarani N, David S, Shamma S. Influence of context and behavior on the population code in primary auditory cortex. J Neurophysiol. 2009; 102:3329–3333. [PubMed: 19759321]

88. Mesgarani N, David SV, Fritz JB, Shamma SA. Mechanisms of noise robust representation of speech in primary auditory cortex. Proc Nat Acad Sci. 2014; 111(18):6792–6797. [PubMed: 24753585]
89. Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF. Reconstructing speech from human auditory cortex. PLOS Biol. 2012; 10(1):e1001251. [PubMed: 22303281]
90. Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012; 485(7397):233–236. [PubMed: 22522927]
91. Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Nat Acad Sci United States Amer. 2012; 109(29):11 854–11 859.
92. O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cerebral Cortex. 2014
93. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Tech Rep. 2001 ITU-T P.862.
94. Nemala S, Zotkin D, Duraiswami R, Elhilali M. Biomimetic multi-resolution analysis for robust speaker recognition. EURASIP J Audio, Speech, Music Process. 2012; 22
95. Carlin MA, Patil K, Nemala SK, Elhilali M. Robust phoneme recognition using biomimetic speech contours. Proc Interspeech. 2012
96. Nemala SK, Patil K, Elhilali M. A multistream feature framework based on bandpass modulation filtering for robust speech recognition. IEEE Trans Audio, Speech, Lang Process. Feb; 2013 21(2): 416–426.
97. Hermansky H, Morgan N. RASTA processing of speech. IEEE Trans Speech Audio Process. Oct; 1994 2(4):382–395.
98. Soltau H, Saon G, Sainath TN. Joint training of convolutional and non-convolutional neural networks. Proc ICASSP. 2014:5572–5576.
99. Meyer BT, Kollmeier B. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Commun. 2011; 53(5):753–767.
100. Patil K, Elhilali M. Goal-oriented auditory scene recognition. Proc Interspeech. 2012
101. Patil K, Elhilali M. Task-driven attentional mechanisms for auditory scene recognition. Proc ICASSP. 2013:828–832.
102. Bellur, A., Elhilali, M. Detection of Speech Tokens in Noise Using Adaptive Spectrotemporal Receptive Fields. Proc. 49th Annu. Conf. Inf. Sci. Syst. (CISS); 2015.

Biographies

Michael A. Carlin (M'05) received the Ph.D. degree in electrical and computer engineering from Johns Hopkins University in 2015. From 2006–2008 he was a Researcher with the Air Force Research Laboratory in Rome, NY, USA. He has also spent summers working with the Human Language Technology Center of Excellence in Baltimore, MD, USA, and MIT Lincoln Laboratory in Lexington, MA, USA. Dr. Carlin is currently a Senior Data Scientist at RedOwl Analytics in Baltimore, MD. His research interests span machine learning,

speech processing, computational neuroscience, and biomimetic approaches for improving noise robustness in speech and audio systems.

Mounya Elhilali (M'00) received her Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 2004. She is now an Assistant Professor at the Department of Electrical and Computer Engineering at the Johns Hopkins University. She is affiliated with the Center for Speech and Language Processing and directs the Laboratory for Computational Audio Perception. Her research examines the neural and computational bases of sound and speech perception in complex acoustic environments; with a focus on robust representation of sensory information in noisy soundscapes, problems of auditory scene analysis and segregation as well as the role of top-down adaptive processes of attention, expectations and contextual information in guiding sound perception. Dr. Elhilali is the recipient of the National Science Foundation CAREER award and the Office of Naval Research Young Investigator award.

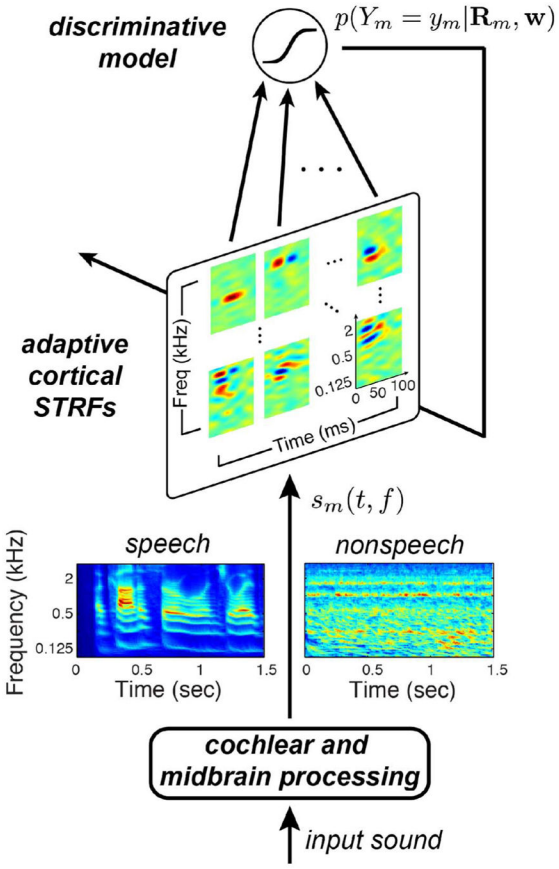

Fig. 1.

Proposed discriminative framework for task-driven STRF adaptation. Examples of speech and non speech stimuli are passed through a model of the auditory periphery, and the resulting auditory spectrogram is analyzed by a bank of STRFs derived from recordings from ferret primary auditory cortex. Top-down feedback acts to assign a behaviorally meaningful categorical label to observed population responses, which are subsequently discriminated using logistic regression. Feedback from the discriminative model, in the form of the regressor prediction error, is used to iteratively adapt the shapes of the STRFs to improve prediction of speech vs. non speech sounds.

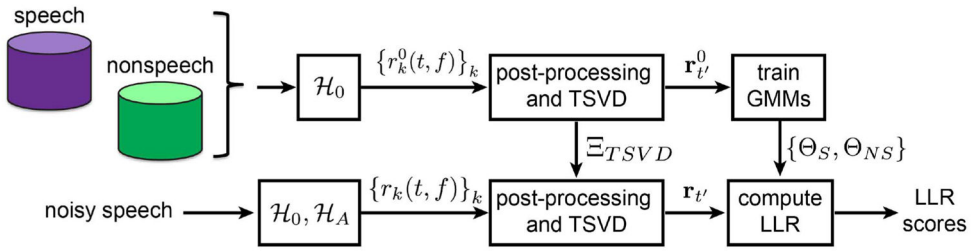

Fig. 2.
Overview of the proposed SAD system.

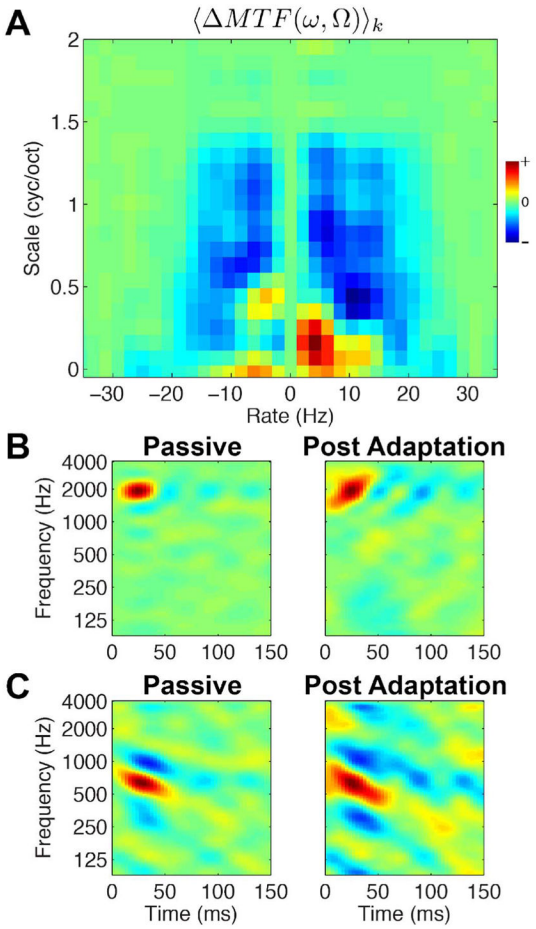

Fig. 3.

Effect of the proposed model for a speech-in-noise detection task. (A) Population effects were quantified by computing the average difference modulation profile $\Delta MTF(\omega, \Omega)$ and show that the model tends to increase sensitivity to slower modulations close to the origin while suppressing faster modulations away from the origin. (B and C) Individual examples illustrating how adaptation causes the STRFs to reorient themselves in a task-driven manner. Results shown for $\lambda = 10^{-1}$, $C = 10^{2.25}$.

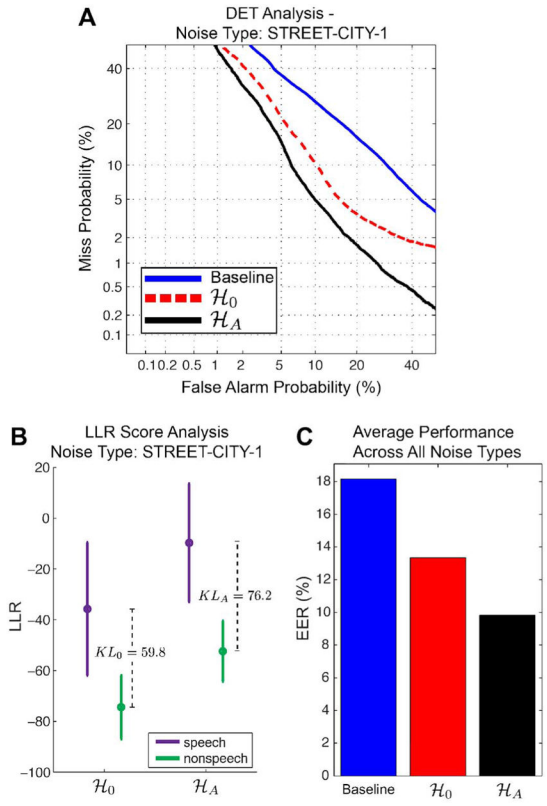

Fig. 4.

SAD Results. (A) DET curves for the street noise condition, computed by pooling scores across all SNRs, for the baseline, passive, and adapted STRFs. (B) Visualization of LLR score distributions (as mean and standard deviation) for the passive and adapted ensembles, showing how use of $\mathcal{H}_A$ improves separation between the speech and nonspeech LLR scores. (C) Average EER over all noise conditions. Adapted STRF ensembles are reported for $\lambda = 10^{-1}$, $C = 10^2$.

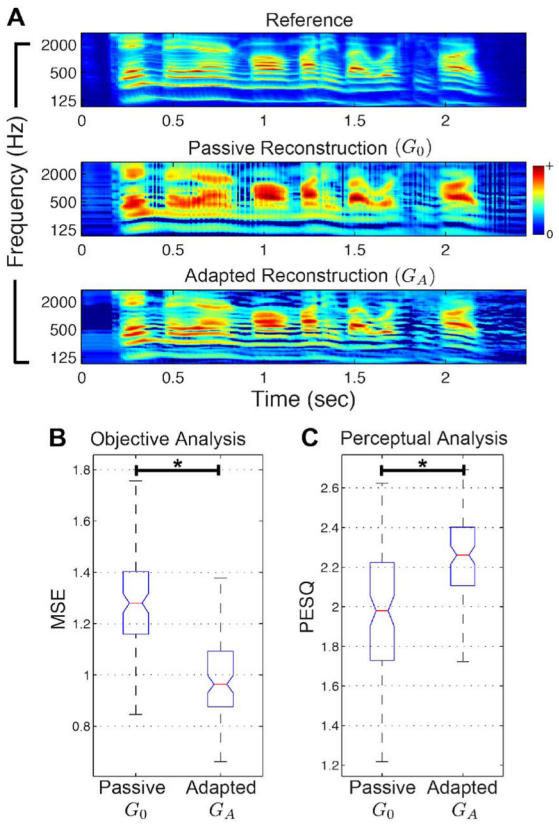

Fig. 5.

Analysis of clean speech reconstructions. (A) A reference spectrogram (top) is compared with reconstructions obtained from the passive (middle) and task-driven STRF ensembles (bottom). (B) Objective analysis shows that G_A yields a significantly better reconstruction compared to G_0. (C) Perceptual analysis shows that G_A yields a significantly higher quality waveform synthesis compared to G_0(*:t-test, $p \approx 0$). Results shown for $\lambda = 10^{-1}$, $C = 10^{2.0}$.

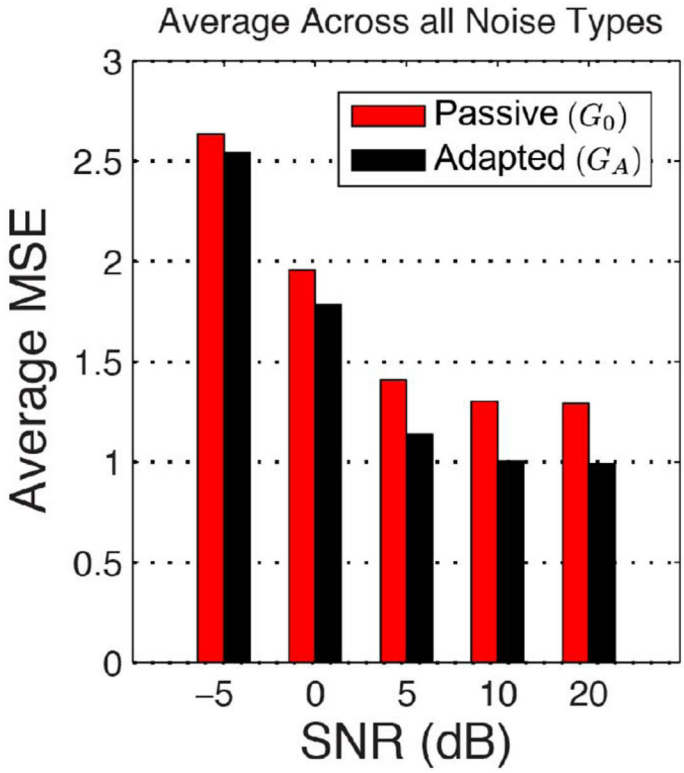

Fig. 6.

Analysis of noisy speech reconstructions. Reconstruction quality degrades with increasing noise. However, in all SNR cases, the adapted ensembles yield, on average, a higher quality reconstruction with respect to the clean references. Results shown for $\lambda = 10^{-1}$, $C = 10^{2.0}$.