Speech processing using adaptive auditory receptive fields

ASHWIN BELLUR AND MOUNYA ELHILALI*

Laboratory for Computational Audio Perception, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

The auditory system exhibits a remarkable ability to adapt to its listening environment, driven both by sensory-based cues and goal-directed processes. Here, we focus on the role of attentional feedback in facilitating processing of speech sounds in presence of nonstationary noises. We examine a theoretical formulation for retuning of cortical-like receptive fields to enable robust detection of speech sounds in presence of interference. The framework employs modulation-tuned filters aimed at emulating tuning characteristics of neurons at the level of auditory cortex. This bank of filters is then modulated based on goal-directed feedback to enhance separability between the feature representation of speech and nonspeech sounds. We hypothesize that this retuning procedure results in an emphasis of unique speech and nonspeech modulations in a high-dimensional space. We discuss the implications of this retuning on the fidelity of encoding speech sounds in presence of seen and novel noise conditions, and discuss implications of such plasticity in facilitating listening in challenging acoustic environments, hence opening the door to adaptive and intelligent audio technology that can emulate the biological system.

INTRODUCTION

When engaged in a conversation in a noisy cafeteria, our brain relies on cognitive processes particularly attention to help navigate the challenging acoustic stimulus impinging on its ears and detect sounds of interest. Attention acts as information bottleneck that sifts through acoustic cues and helps boost the signal-to-noise representation of targets relative to interferers in order to ultimately facilitate processing of these sounds of interest. An increasing body of work suggests that attending to a target sound induces profound but rapid adaptation effects in brain responses. Magnetoencephalography (MEG) recordings in listeners attending to a target speech in presence of competing talkers showed selective enhancement of neural phase-locking to the attended stream resulting in improved and robust reconstruction of the attended speech regardless of the signal-to-noise relative to the interferer (Akram *et al.*, 2016; Ding and Simon, 2012; Puvvada and Simon, 2017). The readout of the attended speech appears to also be in synchrony with enhancement in brain oscillations (particularly alpha rhythm), which selectively modulates the neural representation of the attended stimulus resulting in improved segregation

*Corresponding author: mounya@jhu.edu

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017.
 ISBN: 978-87-990013-6-1.

(Wostmann *et al.*, 2016). Similar results have been reported using electroencephalography (EEG) where selective attention in noisy environments (e.g., competing talkers, reverberation) also improve neural encoding of the speech envelope of the attended stream (Fuglsang *et al.*, 2017; O'Sullivan *et al.*, 2014). A refined look at neural activity at the single neuron level has also corroborated these findings using high-density intracranial electrode arrays in human participants (Mesgarani and Chang, 2012). Results show that neural responses in non-primary auditory cortex (posterior superior and middle temporal gyrus) are driven almost solely by the attended speaker.

A natural question that arises is how does the auditory system balance a stable sensory encoding and perceptual decoding in presence of such profound adaptation effects (Seriès *et al.*, 2009). Given the distributed neural circuitry underlying this attention-induced modulation, one interpretation of these effects is at the perceptual stage whereby adaptation of perceptual estimates implies refining the *interpretation* of sensory encoding for different tasks/environments. This account is often favored in engineering solutions which employ similar forms of adaptation (e.g., domain adaptation, model adaptation) in machine learning to adapt to specific targets or classes or generalize models across conditions of the data (Ben-David *et al.*, 2010; Gauvain and Lee, 1994; Leggetter and Woodland, 1995; Siohan *et al.*, 2001).

An alternative interpretation is that observed effects are in fact due to adaptation of the sensory mapping itself. This form of adaptation implies that cognitive processes might receive inconsistent or suboptimal encoding (Seriès *et al.*, 2009). If feature maps themselves are retuning, they are altering the representation of the incoming stimulus hence requiring perceptual processes to compensate for this warped mapping or at least take it into account. Electrophysiological recordings in single neurons as early as auditory cortex put forth evidence in support of adaptation of sensory feature maps. Cortical activity in animals engaged in various behavioural tasks shows that tuning characteristics of these neurons exhibit rapid tuning shifts in line with the behavioural task at hand (Elhilali *et al.*, 2007; Fritz *et al.*, 2003; Lu *et al.*, 2017; Winkowski *et al.*, 2017). Effects of this adaptation can be gleaned through their neural spectro-temporal receptive fields (STRFs). An STRF is a measure that characterizes the steady state response properties of auditory neurons, spanning their temporal dynamics and spectral selectivity (Elhilali *et al.*, 2013). At the level of auditory cortical areas, these very receptive fields reflect the inherent properties of individual neurons which reshape their tuning to reflect task demands and relevant targets or backgrounds in an auditory scene (Atiani *et al.*, 2014; David *et al.*, 2012; Engineer *et al.*, 2014; Fritz *et al.*, 2005).

In this work, we examine the theoretical underpinnings of the attention-driven receptive field plasticity in shaping neural encoding of incoming sound signals, and effectively enhancing detection of target sounds in complex scenes. We focus this question in the case of listening to speech sounds in presence of noise interferers or distortions such as reverberation. Here, we review recent work which leverages STRF plasticity in models for robust detection of speech in presence of background noise (Bellur and Elhilali, 2017; Carlin and Elhilali, 2015b). We comment on implications

of observed changes from both models in interpreting observed changes in the biological system.

MODELING RECEPTIVE FIELD PLASTICITY

The transformations undertaken along the auditory system can be emulated by a multistage process whereby the incoming acoustic waveform is mapped from a one-dimensional signal representation along time to various feature dimensions that highlight characteristics of the acoustic waveform along both time, frequency, and spectrotemporal modulations. These transformations – achieved through a variety of analysis maps – act as feature detectors to extract cues relevant for processing and interpretation of incoming signals (Eggermont, 2001; Nelken and Bar-Yosef, 2008).

In the current work, we ask the question: How would the system behave if a sensory mapping stage, specifically at the level of cortical processing, would receive feedback that induces changes in its properties in a direction dictated by the feedback signal, and within constrains imposed by the system? We contrast two approaches to achieve such optimization, a linearized vs. nonlinear approach, as discussed next.

A linearized optimization of receptive field plasticity

In a first study, we examine a framework for such feedback defined in a discriminative fashion (Carlin and Elhilali, 2015b). In this setup, the cortical stage is retuned to contrast the mapping of speech and non-speech stimuli. The model starts by transforming all incoming signals into a time-frequency spectrogram, by employing a model of the auditory periphery (Chi *et al.*, 2005). This stage maps the acoustic waveform $x(t)$ through a series of stages including an array of asymmetric, constant-Q band-pass filters, first order derivative, half-wave rectification and spectral derivative, before smoothing the responses using a short time window $w(t,\tau) = e^{-t/\tau}u(t)$ to mimic the loss of phase locking observed at the level of the midbrain. The auditory spectrogram $s(t,f)$ is next processed by an adaptable feature extraction framework, based on the processes of the cortical regions and task-driven plasticity observed in the auditory pathway. Carlin and Elhilali (2015b) propose using an ensemble of adaptable STRFs to extract frequency and spectro-temporal dynamics information from the auditory spectrogram. STRFs used in this work are neurophysiologically-recorded function obtained from non-behaving ferrets (recorded in studies by Elhilali *et al.* (2004) and Fritz *et al.* (2003). These biological STRFs are used as initial spectro-temporal filters upon which attentional feedback will be applied to induce plastic changes in line with the discriminative framework. Since the approach employs biologically-obtained filters in a non-parametric form, it uses a linear model using logistic regression to retune these filters in a manner that enhances the ability of the system to detect speech in a noisy environment.

The adaptive framework is formulated as maximizing the conditional likelihood of labels *y* with respect to the weighted ensemble response *E*, as defined below

$$p(Y = y|\boldsymbol{E},\boldsymbol{w}) \equiv \sigma(y\boldsymbol{w}^T\boldsymbol{E}) \qquad \text{(Eq. 1)}$$

where $\sigma(\gamma) = {}^{1}/_{(1+\exp(-\gamma))}$ is the logistic function and $y \in \{+1, -1\}, y = +1$ denotes speech and $y = -1$ denotes non-speech. Let $r_k(t, f)$ be the firing rate of the k^{th} neuron:

$$r_k(t, f) = h_k(t, f) *_{tf} s(t, f) \quad \text{(Eq. 2)}$$

where $h_k(t, f)$ is the transfer function of the k^{th} STRF and $*_{tf}$ is the 2D convolution over time and frequency axes. The corresponding modulation domain representation can be determined as

$$|R_k(\omega, \Omega)| = |H_k(\omega, \Omega)|.|S(\omega, \Omega)| \quad \text{(Eq. 3)}$$

where $R_k(\omega, \Omega)$, $H_k(\omega, \Omega)$ and $S(\omega, \Omega)$ are the 2D discrete Fourier transforms of the firing rate, STRF and stimulus spectrogram, respectively. ω represents temporal modulations or rates (in Hz) and Ω represents spectral modulations or scale (in cycles/octave). The ensemble response $\boldsymbol{E}$ in Eq. 1 defined as

$$\boldsymbol{E} = [1, \textstyle\sum_{\boldsymbol{\omega\Omega}} |R_1(\omega, \Omega)|, \ldots, \sum_{\boldsymbol{\omega\Omega}} |R_K(\omega, \Omega)|] \in \mathbb{R}^{K+1} \quad \text{(Eq. 4)}$$

is a supervector of responses of the K neurons to a stimulus. $\boldsymbol{w} = [w_0, w_1, \ldots, w_k]$ in equation 1 is the vector of regression coefficients for the K neurons of the ensemble.

Throughout this framework, the model mimics common experimental paradigms whereby neurons are characterized with a 'default' tuning transfer function H_0. These are typically obtained when the auditory system is not engaged in any active task, but is in a passive state. Once the system is engaged in a task, these filter parameters H_0 are retuned, yielding adapted receptive fields H_a. In the proposed framework by Carlin and Elhilali (2015b), the adaptation problem is cast as an optimization with goal to minimize the cost function $J(w, \mathcal{H}_a)$ defined as

$$J(w, \mathcal{H}_a) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \frac{C}{M}\textstyle\sum_m log\big(\sigma(y_m \boldsymbol{w}^T \boldsymbol{E}_m)\big) + \frac{\lambda}{2}\sum_k \|\Delta_k\|_F^2 \quad \text{(Eq. 5)}$$

where $\mathcal{H}_a = \{|H_k^a(\omega, \Omega)|\}_{k=1}^K$ and $\Delta_k = |H_k^a(\omega, \Omega)| - |H_k^0(\omega, \Omega)|$. $H_k^0(\omega, \Omega)$ is the default tuning of the k^{th} neuron and $H_k^a(\omega, \Omega)$ its adapted tuning. By formulating the adaption process in this manner, the framework seeks to obtain a weighted set of retuned neural ensemble that maximizes the conditional probability averaged over all stimuli (M). The Δ_k term ensures that each individual neuron retunes marginally from its default tuning, consistent with the observation that cortical neurons maintain stable properties while adapting marginally to behavioral tasks (Elhilali *et al.*, 2007).

In order to determine the regression parameters $\boldsymbol{w}$ and retuned STRF ensemble $\mathcal{H}_a$, block coordinate descent is employed, alternating between the 2 convex problems

$$argmin\, J(w, \mathcal{H}_a) \quad s.t. \quad |H_k^a(\omega, \Omega)| \geq 0 \; \forall k, \omega, \Omega$$

$$argmin\, J(w, \mathcal{H}_a) \quad s.t. \quad w_k > 0$$

Upon convergence, the solution to these two convex problems can be written as

$$|H_k^a(\omega,\Omega)| = |H_k^0(\omega,\Omega)| + \frac{C}{\lambda}.w_k.\frac{1}{M}\sum_m y_m\left(1-\sigma(y_m \boldsymbol{w}^T \boldsymbol{E}_m)\right)S_m(\omega,\Omega) \quad \text{(Eq. 6)}$$

$$\boldsymbol{w} = \frac{C}{M}\sum_m y_m\left(1-\sigma(y_m \boldsymbol{w}^T \boldsymbol{r}_m)\right)\boldsymbol{r}_m \quad \text{(Eq. 7)}$$

where M denotes the number of stimuli used for the adaptation process.

It can be seen from the constraints and solution equations (Eqs. 6 and 7) that by enforcing the weights to be positive and using the labels $y_m = +1$ for speech and $y_m = -1$ for non-speech, the adaptation process seeks to enhance speech modulation while suppressing non-speech content. Another interesting observation relates to the impact of the stimulus. By interpreting $1-\sigma(y_m \boldsymbol{w}^T \boldsymbol{E}_m)$ as prediction error, certain stimuli that are too difficult to predict have a stronger impact on the adaptation process. Furthermore, it can be seen in Eq. 7 that neurons that are task-relevant receive larger weights in contrast to the task-irrelevant neurons.

A nonlinear parametric optimization of receptive field plasticity

In contrast to the approach described above, Bellur and Elhilali (2017) explore an alternate framework to model task-driven plasticity, focusing on 3 broad different takes to the optimization problem: First, the approach in Bellur and Elhilali (2017) employs parameterized Gabor filters to encode spectrotemporal dynamics, instead of physiologically recorded receptive fields. By employing parameteric functions to emulate cortical receptive fields, Gabor filters can be re-tuned to achieve a non-linear transformation in contrast to the linear adaptation of filter patches as used in Carlin and Elhilali (2015b). Second, instead of assigning fixed class labels $y_m = \pm 1$ to distinguish speech from non-speech tokens, the approach in Bellur and Elhilali (2017) employs a generative probabilistic model using Gaussian mixture models (GMMs) to serve as *object* representations of clean speech and non-speech classes (Duda *et al.*, 2000). In this case, the optimization seeks to retune the Gabor filters in a manner that enhances the ability of the GMMs to discriminate between noisy speech and nonspeech, thereby adapting the feature extraction process to work even under novel noise conditions. Third, the optimization process employs a Genetic algorithm (Michalewicz, 1996). This approach differs from the convex optimization formulated in Carlin and Elhilali (2015b) and allows to search the parameter space for the Gabor filters to ensure improved discrimination between the two classes with respect to the fixed GMMs.

This approach follows the same general framework as presented earlier. A time-domain waveform is first mapped through a model of the auditory periphery to derive an auditory spectrogram $s(t,f)$. Then, a bank of 2D Gabor filters are applied to analyze the spectral and temporal modulations in the spectrogram. Such filters are considered a reasonable approximation of cortical receptive fields observed in the mammalian auditory system (Ezzat *et al.*, 2007; Theunissen *et al.*, 2000). The filters are parametrized as:

$$g_k(t,f) = \frac{\alpha_k}{2\pi\sigma_{tk}\sigma_{fk}} e^{-\frac{1}{2}\left(\frac{t_1^2}{\sigma_{tk}^2}+\frac{f_1^2}{\sigma_{fk}^2}\right)} e^{2\pi j(\omega_k t+\Omega_k f)} \quad \text{(Eq. 8)}$$

where $t_1 = tcos(\theta_k) + fsin(\theta_k)$ and $f_1 = -tsin(\theta_k) + fcos(\theta_k)$. σ_{tk} and σ_{fk} denote the temporal and spectral bandwidths of the Gaussians of the k^{th} Gabor filter, respectively. θ_k specifies the orientation of the main lobe of the Gabor filter and α_k is a gain term. ω_k and Ω_k are the rate and scale of the k^{th} Gabor filter.

The auditory spectrogram is convolved with a bank of Gabor filters $\mathcal{g} = \{g_1, g_2, \dots, g_K\}$ spanning the spectrotemporal space set by the chosen parameters (Eq. 9). The output is then collapsed along the time axis to obtain the spectrotemporal dynamics and frequency information as shown in equation Eq. 10.

$$C_k(t,f) = |s(t,f) *_{tf} g_k(t,f)| \quad \text{(Eq. 9)}$$

$$T_k(f) = \int C_k(t,f)dt \quad \text{(Eq. 10)}$$

Like the regression approach, the Gabor filter model in Bellur and Elhilali (2017) starts with a default set of parameters $\mathcal{g}^0 = \{g_1^0, g_2^0, \dots, g_K^0\}$ analogous to the passive receptive fields used in Carlin and Elhilali (2015b). The Gabor parameters are then retuned for robust speech activity detection, to obtain an adapted filter bank of Gabor filters denoted as $\mathcal{g}^a = \{g_1^a, g_2^a, \dots, g_K^a\}$. These adapted filters are derived based on statistical models of speech and non-speech data; Gaussian mixture models of clean speech and nonspeech estimated based on their spectrotemporal modulation features (Eq. 10). A held out set of noisy speech and nonspeech data is then used adapt the filters in manner that enhances the ability of the GMMs to discriminate between noisy speech and nonspeech even in mismatched conditions. The hypothesis at the center of this work is that this retuning process will lead to highlighting the *discriminable regions* of the spectrotemporal modulation space as represented by the GMMs, hence resulting in robust speech activity detection under novel noise conditions.

The Gabor filters are retuned using a genetic algorithm which scans the parameter space. It employs a fitness measure to gauge the suitability of the parameter choice. In Bellur and Elhilali (2017), the fitness measure used is d-prime, defined as:

$$d' = \frac{\mu_{ns} - \mu_s}{\sqrt{\frac{1}{2}(\sigma_{ns}^2 + \sigma_s^2)}} \quad \text{(Eq. 11)}$$

μ_c and σ_c denote the mean and standard deviation respectively of the log likelihood ratio (LLR) values estimated using the GMMs trained on clean speech (c=s) and nonspeech (c=ns) data. The genetic algorithm is initialized with the default parameters ($\mathcal{g}^0$) as a member of the first generation. The algorithm then propagates through multiple generations to find the fittest member ($\mathcal{g}^a$) as defined by the equation Eq. 11.

OPTIMIZED MAPPING OF SPEECH AND NONSPEECH SOUND CLASSES

Figure 1A shows results of the adaptation process in terms of the average difference between the modulation profiles of the STRF after and before adaptation; That is $\langle |H_k^a(\omega,\Omega)| - |H_k^0(\omega,\Omega)| \rangle_k$ where $\langle . \rangle_k$ denotes averaging. The figure illustrates that the neural ensemble tends to emphasize slower modulations especially for positive rates (which correspond to downward modulations), which are commensurate with modulations in speech sounds (Elliott and Theunissen, 2009). Given the choice of

label values and the fact that the weights $\boldsymbol{w}$ are set to be positive, the adaptation framework also leads to suppression of responses to faster modulations, hence diminishing the response to non-speech regions of the spectrotemporal space.

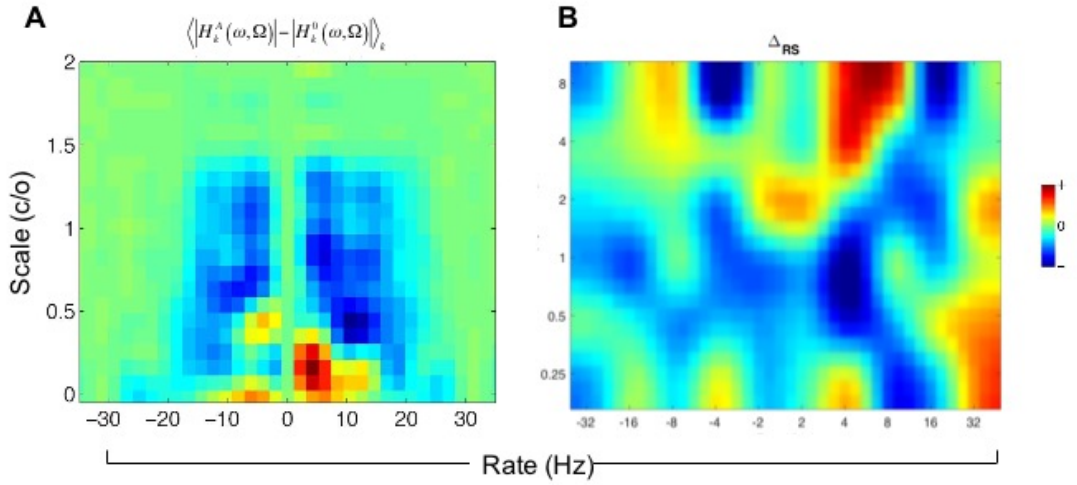

Fig. 1: (A) Average difference in the responses of STRFs before and after adaptation using linearized regression. The difference is measured as $\langle |H_k^a(\omega,\Omega)| - |H_k^0(\omega,\Omega)| \rangle_k$ where $\langle . \rangle_k$ denotes the average operation [Figure reproduced from (Carlin and Elhilali, 2015a) with permission from IEEE]. **(B)** Δ_{RS} difference between the energies in the rate scale space on using $\mathcal{g}^a$ and $\mathcal{g}^0$ filter banks in the nonlinear optimization approach using Gabor filters.

Figure 1B shows the difference between the energies in the rate scale space on using $\mathcal{g}^a$ and $\mathcal{g}^0$ filter banks. Δ_{RS} depicted in this figure is estimated as:

$$\Delta_{RS} = \langle \sum_f \sum_t |s_m(t,f) *_{tf} \mathcal{g}^a| - \sum_f \sum_t |s_m(t,f) *_{tf} \mathcal{g}^0| \rangle_m \qquad \text{(Eq. 12)}$$

where $\langle . \rangle$ denotes the average over all stimuli, both noisy speech and nonspeech. Δ_{RS} illustrates the difference in energies on projecting the stimuli on to the spectrotemporal modulation space using the 2 sets of Gabor filter banks. It can be seen that while slower modulations are emphasized, broadband fast modulations are also emphasized, as well fast spectral modulation at 4-Hz rate. The figure also suggests that greater discriminability is attained on adapting the filters because sparse non-overlapping regions of speech and nonspeech are emphasized on adaptation, while overlapping regions are suppressed.

Further insight into the behavior of the Gabor model can be gleaned from contrasting the log-likelihood estimates with respect to both speech and non-speech data. Figure 2A shows the histogram of the log likelihood ratio values of noisy speech and non-speech stimuli estimated, before ($\mathcal{g}^0$) and after adaptation ($\mathcal{g}^a$) of the Gabor filters. As can be seen from the plots, the classes are more separable on using the retuned

filter bank. It is interesting to note that on adaptation, the LLR values for the 2 classes do not necessarily move in the opposite directions, rather they become narrower owing to the fact that the *d*-prime measures reward lesser spread of the LLR values for a class. Figures 2B and 2C show a schematic summarizing the impact of the different optimization approaches on the resulting representation of speech and nonspeech classes.

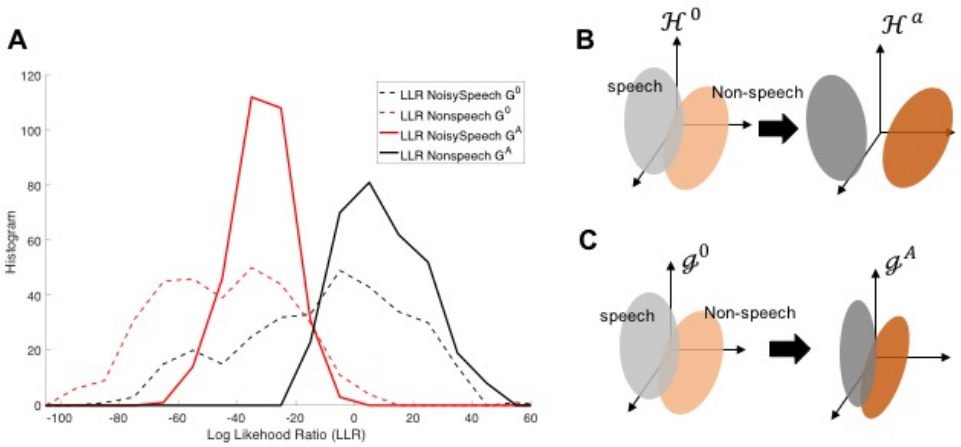

Fig. 2: (A) Histogram of the log likelihood ratio values of noisy speech and nonspeech stimuli before (g^0) and after adaptation (g^a) of the Gabor filters. **(B, C)** Schematic of changes in mapping of speech and non-speech classes using linearized regression vs. nonlinear optimization.

CONCLUSIONS

The models reviewed here shed light on two possible strategies that improve speech detection in noise: (i) An approach that pushes the perceptual maps of speech and nonspeech further apart from each other (Fig. 2B). This is achieved by reshaping the feature maps to emphasize acoustic cues unique to speech and de-emphasize characteristics of nonspeech. As shown in Fig. 1A, putting more emphasis on slow temporal modulations in the region around ~4 Hz results in highlighting areas known to correlate well with characteristics of speech signals (e.g., syllabic rate, Elliott and Theunissen, 2009). This outcome is achieved through a linearized optimization of cortical receptive fields that allows minor tweaks to their response properties in a linear way.

In contrast, a parametrized approach that exhaustively searches the space of cortical filters represented as Gabor functions combined with statistical modeling of the perceptual decision space results in a different outcome by tightening the perceptual maps of speech vs. nonspeech classes (Fig. 2C). This outcome is an equally acceptable solution to the stated problem, and in fact has been shown to yield superior performance of speech detection in noise, especially when contrasted with novel noisy speech and nonspeech conditions.

Overall, either strategy (or combined) offers a robust biomimetic approach to adaptive signal processing to improve sound perception in noise. It remains to be seen which approach is more in line with scheme underlying neural plasticity in the brain. As more advanced experimental techniques emerge and paradigms are able to train animals on more sophisticated behavioral tasks, it will be possible to tease apart the theoretical underpinnings of attention-driven neural plasticity in the auditory system.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health under grant R01HL133043 and the Office of Naval research under grant ONR N000141612045.

REFERENCES

Akram, S., Presacco, A., Simon, J.Z., Shamma, S.A., and Babadi, B. (**2016**). "Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling," Neuroimage, **124**, 906-917. doi: 10.1016/j.neuroimage.2015.09.048

Atiani, S., David, S.V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S.A., *et al.* (**2014**). "Emergent selectivity for task-relevant stimuli in higher-order auditory cortex," Neuron, **82**, 486-499. doi: 10.1016/j.neuron.2014.02.029

Bellur, A., and Elhilali, M. (**2017**). "Feedback-driven sensory mapping adaptation for robust speech activity detection," IEEE T. Audio Speech, **25**, 481-492. doi: 10.1109/TASLP.2016.2639322

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J.W. (**2010**). "A theory of learning from different domains," Mach. Learn., **79**, 151-175. doi: 10.1007/s10994-009-5152-4

Carlin, M.A., and Elhilali, M. (**2015a**). "A framework for speech activity detection using adaptive auditory receptive fields," IEEE T. Audio Speech, **23**, 2422-2433. doi: 10.1109/TASLP.2015.2481179

Carlin, M.A., and Elhilali, M. (**2015b**). "Modeling attention-driven plasticity in auditory cortical receptive fields," Front. Comput. Neurosci., **9**, 106. doi: 10.3389/fncom.2015.00106

Chi, T., Ru, P., and Shamma, S.A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am., **118**, 887-906.

David, S.V., Fritz, J.B., and Shamma, S.A. (**2012**). "Task reward structure shapes rapid receptive field plasticity in auditory cortex," Proc. Natl. Acad. Sci. USA, **109**, 2144-2149. doi: 10.1073/pnas.1117717109

Ding, N., and Simon, J.Z. (**2012**). "Emergence of neural encoding of auditory objects while listening to competing speakers," Proc. Natl. Acad. Sci. USA, **109**, 11854-11859. doi: 10.1073/pnas.1205381109

Duda, R.O., Hart, P.E., and Stork, D.G. (**2000**). *Pattern Classification.* Wiley.

Eggermont, J.J. (**2001**). "Between sound and perception: reviewing the search for a neural code," Hear. Res., **157**, 1-42.

Elhilali, M., Fritz, J.B., Klein, D.J., Simon, J.Z., and Shamma, S.A. (**2004**). "Dynamics of precise spike timing in primary auditory cortex," J. Neurosci., **24**, 1159-1172. doi: 10.1523/JNEUROSCI.3825-03.2004

Elhilali, M., Fritz, J.B., Chi, T.-S., and Shamma, S.A. (**2007**). "Auditory cortical receptive fields: Stable entities with plastic abilities," J. Neurosci., **27**, 10372-10382. doi: 10.1523/JNEUROSCI.1462-07.2007

Elhilali, M., Shamma, S.A., Simon, J.Z., and Fritz, J.B. (**2013**). "A linear systems view to the concept of STRF," in *Handbook of Modern Techniques in Auditory Cortex*. Eds. D. Depireux and M. Elhilali (Nova Science Pub Inc), 33-60.

Elliott, T.M., and Theunissen, F.E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS Comput. Biol., **5**, e1000302.

Engineer, C.T., Perez, C.A., Carraway, R.S., Chang, K.Q., Roland, J.L., and Kilgard, M.P. (**2014**). "Speech training alters tone frequency tuning in rat primary auditory cortex," Behav. Brain Res., **258**, 166-178. doi: 10.1016/j.bbr.2013.10.021

Ezzat, T., Bouvrie, J.V, and Poggio, T. (**2007**). "Spectro-temporal analysis of speech using 2-d Gabor filters," Proc. Interspeech, 506-509.

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (**2003**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nat. Neurosci., **6**, 1216-1223. doi: 10.1038/nn1141

Fritz, J.B., Elhilali, M., and Shamma, S.A. (**2005**). "Rapid task-dependent plasticity in primary auditory cortex," in *Auditory Cortex-Towards a Synthesis of Human and Animal Research*. Wds. P. Heil, R. Konig, E. Budinger, and H. Scheich (Mahwah, NJ: Lawrence Erlbaum Associates), 445-466.

Fuglsang, S.A., Dau, T., and Hjortkjær, J. (**2017**). "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," Neuroimage, **156**, 435-444. doi: 10.1016/j.neuroimage.2017.04.026

Gauvain, J.-L., and Lee, C.-H. (**1994**). "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE T. Speech Audio, **2**, 291-298.

Leggetter, C.J., and Woodland, P.C. (**1995**). "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Comput. Speech Lang., **9**, 171-185.

Lu, K., Xu, Y., Yin, P., Oxenham, A.J., Fritz, J.B., and Shamma, S.A. (**2017**). "Temporal coherence structure rapidly shapes neuronal interactions," Nat. Commun., **8**, 13900. doi: 10.1038/ncomms13900

Mesgarani, N., and Chang, E.F. (**2012**). "Selective cortical representation of attended speaker in multi-talker speech perception," Nature, **485**, 233-236. doi: 10.1038/nature11020

Michalewicz, Z. (**1996**). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer Science & Business Media.

Nelken, I., and Bar-Yosef, O. (**2008**). "Neurons and objects: The case of auditory cortex," Front. Neurosci., **2**, 107-113. doi: 10.3389/neuro.01.009.2008

O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., *et al.* (**2014**). "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," Cereb. Cortex., 1697-1706. doi: 10.1093/cercor/bht355

Puvvada, K.C., and Simon, J.Z. (**2017**). "Cortical representations of speech in a multitalker auditory scene," J. Neurosci., **37**, 9189-9196. doi: 10.1523/JNEUROSCI.0938-17.2017

Seriès, P., Stocker, A.A., and Simoncelli, E.P. (**2009**). "Is the homunculus "aware" of sensory adaptation?" Neural Comput., **21**, 3271-3304.

Siohan, O., Chesta, C., and Lee, C.-H. (2001). "Joint maximum a posteriori adaptation of transformation and HMM parameters," IEEE T. Speech Audio, **9**, 417-428. doi: 10.1109/89.917687

Theunissen, F.E., Sen, K., and Doupe, A.J. (**2000**). "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," J. Neurosci., **20**, 2315-2331.

Winkowski, D.E., Nagode, D.A., Donaldson, K.J., Yin, P., Shamma, S.A., Fritz, J.B., *et al.* (**2017**). "Orbitofrontal cortex neurons respond to sound and activate primary auditory cortex neurons," Cereb. Cortex, 1-12. doi: 10.1093/cercor/bhw409

Wostmann, M., Herrmann, B., Maess, B., and Obleser, J. (**2016**). "Spatiotemporal dynamics of auditory attention synchronize with speech," Proc. Natl. Acad. Sci. USA, **113**, 3873-3878. doi: 10.1073/pnas.1523357113