Detection of speech tokens in noise using adaptive spectrotemporal receptive fields

Ashwin Bellur
Department of Electrical and
Computer Engineering
Johns Hopkins University
Email: abellur1@jhu.edu

Mounya Elhilali
Department of Electrical and
Computer Engineering
Johns Hopkins University
Email: mounya@jhu.edu

***Abstract*—Neurophysiological studies of sound encoding at the level of auditory cortex paint a picture of an intricate filterbank that encodes detailed spectral and temporal modulations in the sensory input. Furthermore, these filters exhibit adaptive qualities called neural plasticity that shape their tuning parameters in line with behavioral goals of interest. In this work, we explore qualitative principles about how this neuronal reshaping can aid in an enhanced representation of target sounds. Here, we employ a set of parameterized two-dimensional Gabor filters as basis functions that tile the space of neurophysiological spectrotemporal modulations. We examine mechanisms for judiciously retuning parameters of the Gabor filterbank in order to enhance the representation of target sounds of interest. We test the efficacy of this scheme in enhancing representation of sound tokens in adverse noisy backgrounds.**

I. Introduction

Everyday acoustic environments are complex sensory signals that are composed of multiple sound "objects". While the identity of the sound "object" is not necessarily uniquely constructed [1], sounds of interest are often determined by the listener in line with their behavioral goals. When chatting at a cocktail party, a listener is naturally trying to attend to the speech signal produced by an interlocutor; hence focusing their attention on it as their object of interest, all the while ignoring surrounding interference or background chatter.

Parsing such acoustic scenes is naturally challenging given the complex temporal and spectral dynamics of the sound that originates from the multiple sources present in the acoustic scene. Despite the complex nature of the signal, humans exhibit an effortless ability to parse such complex scenes and robustly encode signals of interest even amidst severe interference. Studies [2] of the mammalian auditory pathway have shown that neurons in the central auditory system particularly in primary auditory cortex play a major role in parsing such complex acoustic scenes [3], [4]. Cortical neurons and their corresponding transfer functions called spectrotemporal receptive fields [5] act as selective filters that encode the details of the temporal dynamics (or temporal modulations) and frequency changes (or spectral modulations) in a signal. This mapping effectively projects a low dimensional acoustic waveform onto a high dimensional modulation space that facilitates the separation of different auditory objects present in a scene. This behavior can be nicely modeled using modulation selective filters, such as 2-dimensional Gabor functions. While Gabor filters are only a linear approximation of the complex selectivity in cortical neurons, they represent a good approximation of the basis functions that tile the modulation space observed in neurophysiology. Moreover, a number of studies have in fact shown that such Gabor approximations of cortical tuning can be quite effective in extracting joint spectrotemporal modulations from the signal [6], [7].

In addition to their intricate selectivity, cortical neurons also exhibit the ability to adapt their tuning properties to behavioral and task demands as dictated by the attentional state of the brain. These changes referred to as task-driven plasticity are reported as changes in the shape and/or gain of the cortical receptive fields [8], [9]. These changes have been argued to operate as a contrast matched filter so as to highlight a target object of interest while suppressing undesired background distractors [10]. The net effect of these changes is that the representation of a target sound is effectively enhanced relative to the background; hence facilitating target detection. Figure 1 shows a reproduction of an example of STRF plasticity observed in neurophysiological recordings [11]. In this figure, a cortical STRF (left panel) is recorded while an awake ferret is passively listening to modulated noise sounds. The figure shows tuning properties of that particular neuron, particularly its strong inhibitory field around 9KHz. The animal is then engaged in an active listening task, where it has to detect the presence of a pure tone at 7KHz that is presented intermittently with the modulated noise sounds. The right panel shows the recorded STRF from the same neuron during active detection. The filter reveals an increased excitatory response around 7KHz, hence enhancing the neuron's ability to detect the presence of the desired tone.

In the current study, we examine the role of such plasticity retuning in facilitating detection of speech tokens in presence of background noise. In order to closely dissect the role of different components of the receptive field in improving the detection of the target sound, we employ a set of 2-dimensional Gabor filters as linear approximations of the observed neurophysiological STRFs. The benefit of working with parameterized filters is that we can independently manipulate their individual components and assess their contribution to the improved detection of the target. This work examines mechanisms for judiciously retuning parameters of the Gabor filterbank. This manipulation is performed in the context of speech detection and focuses on the role of different filter parameters in the detection of specific speech phonemes.

978-1-4799-8428-2/15/$31.00 ©2015 IEEE

II. CORTICAL MAPPING OF SPEECH SIGNALS

The transformation of an acoustic signal at the level of the mammalian auditory system can be structured into two basic stages. In the early subcortical processing stage, the acoustic signal is transformed into a time frequency representation referred to as the auditory spectrogram [12]. This stage consists of cochlear filtering performed via a set of 128 asymmetric filters spanning 5.3 octaves starting at a frequency of $180Hz$. The resultant signal is then spectrally enhanced using a spectral derivative and a half wave rectification, modeling the lateral inhibition network at the level of the cochlear nucleus. Then, the signal undergoes short term integration with a window $w(t;\tau) = e^{-t/\tau}u(t)$ where $\tau = 8ms$ and a cubic root compression mimicking midbrain processing. If $a(t)$ is the acoustic signal and $h_c(t,f)$ represents the impulse response of the cochlear filters, the subcortical transformation can be written as shown in equation 1. The symbol $*$ denotes convolution with respect to time.

$$Y(t,f) = (max(\delta_f(a(t) * h_c(t,f)), 0) * w(t,\tau))^{1/3} \quad (1)$$

The second stage of processing is the cortical stage where the estimated auditory spectrogram undergoes further spectrotemporal analysis. This stage mimics the modulation analysis believed to take place at the level of primary auditory cortex. Response fields of cortical neurons known as spectrotemporal receptive fields (STRF) can be computationally modeled as a bank of modulation selective filters with spectrotemporal impulse responses $S(t,f)$. The resultant cortical output $r(t,f)$ parameterized by frequency f is then given by:

$$r(t,f) = \int_{\tau} S(t,f)Y(t-\tau,f)d\tau \quad (2)$$

In the current work, we approximate the cortical receptive fields using two dimensional Gabor functions. Each of the Gabor filters is tuned to a specific temporal modulation referred to as rate and spectral modulation referred to as scale. The bank of modulation selective Gabor filters are derived using Equation 3.

$$G(t,f,\omega,\Omega|\lambda) = \frac{\alpha_{\omega\Omega}}{2\pi\sigma_{t_{\omega\Omega}}\sigma_{f_{\omega\Omega}}} e^{-\frac{1}{2}(\frac{t_1^2}{\sigma_{t_{\omega\Omega}}^2}+\frac{f_1^2}{\sigma_{f_{\omega\Omega}}^2})} e^{2\pi i(\omega t+\Omega f)} \quad (3)$$

where $t_1 = tcos(\theta_{\omega\Omega}) + fsin(\theta_{\omega\Omega})$ and $f_1 = -tsin(\theta_{\omega\Omega}) + fcos(\theta_{\omega\Omega})$.

The parameters involved in computings the Gabor filters are:

- ω: in Hz is the specific rate of temporal modulations and Ω: in cycles/octave is the scale of spectral modulations. In accordance with neurophysiological findings, the rates (ω) values used to design the filter bank, ranges from 2 to 32 Hz and scale (Ω) ranges from 0.25 to 8 cycles/octave. The Gabor filters can be downward or upward selective. The upward selective filters are denoted using negative rate values.
- $\sigma_{t_{\omega\Omega}}$ and $\sigma_{f_{\omega\Omega}}$ denote the bandwidths of the gaussians of the Gabor filters along time and frequency direction respectively. The initial spread of all the filters are tuned in such way as to include 2 cycles of the sinusoid both along time and frequency axis.
- $\theta_{\omega\Omega}$ specifies the orientation of the main lobe of the Gabor filters. The orientation is specified in degrees and default value of $\theta_{\omega\Omega}$ is set to zero for all the filters.
- $\alpha_{\omega\Omega}$ is an additional gain term used in this work. The scalar gain term can be used to suppress or enhance the output of the filter. As initial value, $\alpha_{\omega\Omega}$ is set to one for all filters.
- Symbol λ collectively represents these four parameters, that is $\lambda = (\sigma_{t_{\omega\Omega}}, \sigma_{f_{\omega\Omega}}, \theta_{\omega\Omega}, \alpha_{\omega\Omega})$. As can be seen from Equation 3, the parameters are indexed by the rate and scale values of the particular filter.

The bank of Gabor filters are then convolved with the auditory spectrogram to obtain the high dimensional representation. In this work, we perform a two dimensional convolution as shown in Equation 4 to obtain a four dimensional tensor representation.

$$R(t,f,\omega,\Omega) = |Y(t,f) \otimes G(f,t;\omega,\Omega)| \quad (4)$$

When dealing with relatively stationary sounds, a three dimensions rate-scale frequency (RSF) representation, computed using Equation 5 can also be a useful representation of the acoustic signal while abstracting the fine details of the temporal signal itself. In this representation, the temporal dynamics are only captured using the rate axis which represents the relatively slow temporal modulations in the signal. Further dimensionality reduction can also be obtained by collapsing over the frequency dimension, hence resulting in a rate-scale representation which details the energy spread at different rates and scales.

$$T(f,\omega,\Omega) = \int R(t,f,\omega,\Omega)dt \quad (5)$$

III. ADAPTATION OF CORTICAL FILTERS

The two stages of auditory processing discussed map the acoustic signal to a high dimensional feature space capturing a rich representation of the acoustic events. Neurophysiological findings have shown that during attention, in a task drive setting, this high dimensional representation is complemented with cognitive mechanisms that further enhance our ability to identify and comprehend sound events of interest in the presence of other background sound events. The cortical STRFs in fact adaptively reshape [8], [10], [11], in order to accentuate the overall representation of the sound event of interest while suppressing the background sound.

Figures 1 and 2 show reproduction of couple of examples of plasticity observed in ferrets [11]. Figure 1 shows the STRF plasticity observed in ferrets when attempting to detect a tone in the presence of modulated noise sounds. As was discussed in Section I, during attention (Figure 1b), there is an increased excitatory response at the frequency of the tone, improving the ability of the ferret in detecting the tone.

Figure 2 shows another example of plasticity of STRF observed in ferrets. In this case, the ferret is trying to discriminate between two tones. The STRF illustrated in this case is

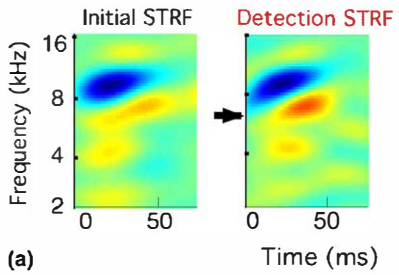

Fig. 1. STRF plasticity during tone detection. Red regions indicate excitation while blue regions indicate inhibition. Image reproduced from [11]

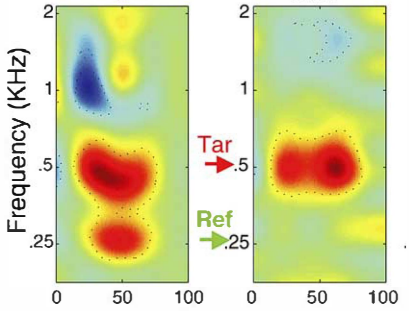

Fig. 2. STRF Plasticity during tone discrimination. Increased excitatory response is observed at the target tone while the response at the reference tone is suppressed. Image reproduced from [11]

originally tuned to $250Hz$ and $500Hz$. During behavior, the STRF reshapes in a manner that leaves the excitatory field at the target as is, while suppressing the reference tone. This contrast matched type of behavior of the STRF plasticity aids in enhancing the performance of the animal in discriminating between the tones.

In this work, we explore modeling the plasticity behavior by retuning the parameters of the Gabor filters. We claim that in a task driven setting, by altering the values of the Gabor parameters $\lambda = (\sigma_{t_{\omega\Omega}}, \sigma_{f_{\omega\Omega}}, \theta_{\omega\Omega}, \alpha_{\omega\Omega})$ as used in Equation 3, one can achieve the desired objective of plasticity, that is, focusing the spotlight on the target sound in the presence of background noise.

In order to illustrate the flexibility offered by the proposed system, we set up the task as follows. Let $a_c(t)$ be the clean target signal and and $Y_c(t,f)$ be the corresponding auditory spectrogram estimated using Equation 1. Let $a_n(t)$ be the target speech signal plus additive noise, with $Y_n(t,f)$ being its corresponding auditory spectrogram representation. The high dimension tensor representation are R_c and R_n respectively obtained on convolving with bank of Gabor filters $G(t,f,\omega,\omega|\lambda)$ as shown in Equations 6 and 7.

$$R_c(t,f,\omega,\Omega) = |Y_c(t,f) \otimes G(t,f,\omega,\Omega|\lambda)| \quad (6)$$
$$R_n(r,f,\omega,\Omega) = |Y_n(t,f) \otimes G(t,f,\omega,\Omega|\lambda)| \quad (7)$$

In a task driven setting, the set of parameters λ are retuned, marginally from their default values. We denote the new set of parameters as $\hat{\lambda}$ and the tensor representation obtained using the altered set of filters (Equation 8) as $\hat{R_n}(t,f,\omega,\Omega)$.

$$\hat{R_n}(t,f,\omega,\Omega) = |Y_n(t,f) \otimes \hat{G}(t,f,\omega,\Omega|\hat{\lambda})| \quad (8)$$

This desired behavior of such an adaptation process would be to obtain $\hat{R_n}(t,f,\omega,\Omega)$ such that $f(R_c,\hat{R_n}) > f(R_c,R_n)$, where function $f()$ is a measure of similarity. Given that relatively stationary additive noise types are being used as examples in this work, one can equivalently measure $f(T_c,\hat{T_n})$ and $f(T_c,T_n)$, where T_c, T_n and $\hat{T_n}$ are the rate-scale frequency (RSF) representation obtained using Equation 5. In this work, we use cosine similarity between RSF representations as a measure of effectiveness of the proposed method. Given that the rate-scale-frequency space is nonnegative, the similarity measure neatly falls in the range [0 1].

IV. Adaptive Detection of Speech Tokens

In this section, we study the effectiveness of the proposed attention modeling mechanism in the context of detecting speech tokes in adverse noisy conditions. Using a repertoire of speech token in noise examples, we illustrate the benefits of retuning the gain ($\alpha_{\omega\Omega}$), bandwidth parameters ($\sigma_{t_{\omega\Omega}}, \sigma_{f_{\omega\Omega}}$) and orientation ($\theta_{\omega\Omega}$). We will illustrate that upon marginally retuning the parameters based on the category of the speech token and the noise conditions, a considerable improvement in detection of speech tokens can be attained. The speech tokens were generated using the PRAAT synthesizer [13]. Different kinds of noise from the Noisex database [14] are used as additive noise in these examples.

A. Gain adaptation

The term $\alpha_{\omega\Omega}$ term introduced in Equation 3 is a scalar multiple that can be used to either emphasize or diminish the response of a Gabor filter tuned rate ω and scale Ω. This gain term assists in enforcing prior knowledge about the target class in the modulation space. Similar gain based approaches have been shown to be useful in identifying the auditory scene [15] or denoising noisy speech signals [16]. The gain term is especially useful in cases where the target and the background are clearly separable in the modulation space. Instead of the default value $\alpha_{\omega\Omega} = 1; \forall\omega,\Omega$, the gain values can be retuned such that, $\alpha_{\omega\Omega} > 1$ can be used for those Gabor filters whose modulation sensitivity (ω,Ω) coincides with that of the target. Similarly gain can be retuned to $\alpha_{\omega\Omega} < 1$ for those Gabor filters at whose rates and scales the noise is predominant.

To illustrate the benefits of retuning gain, we use vowels in noise as examples. While cosine similarity measures are estimated using rate-scale-frequency (RSF) representations, for the sake of illustration, we use the 2 dimensional rate-scale representation of the signal, which is the tensor representation collapsed across both time and frequency. Figure 3a shows the rate-scale representation of the clean vowel *ah* and Figure 3b shows the rate scale representation after adding white noise at an SNR of 0dB. A gain of $\alpha_{\omega\Omega} = 1$ was used $\forall\omega,\Omega$ in obtaining these two representations. As can be seen from both these figures, the target and noise are fairly separable in this high dimensional space. While for the clean target, lower rates

and scales dominate, adding white noise introduces energy at higher rates and scale corrupting the vowel representation.

We propose that by increasing the gain at regions where the vowel is dominant and deemphasizing the regions where noise is dominant, one can achieve a denoised representation of the noisy vowel *ah* shown in Figure 3d. The appropriate choice of retuned gain values $\hat{\alpha}_{\omega\Omega}$ are shown in Figure 3c. The retuned gain values range from 0.6 to 1.4 as indicated by the colorbar. A cosine similarity measure of 0.8712 was estimated between the RSF representation of the clean and the noisy vowel on using the default bank of Gabor filters. On re-estimating the RSF representation of noisy vowel using the adapted filters a cosine similarity measure improved to a value of 0.9481.

Figure 4 shows a similar effect on adapting the gains for Vowel *o* in the presence of white noise at an SNR of -5dB. In this case the similarity measures improved from 0.8212 to 0.9131 on modifying the gain of the filters as indicated in Figure 4c. These two examples indicate that for fairly separable signals in the modulation space, retuning the gain suffices in improving the ability to detect the target even in adverse noise conditions.

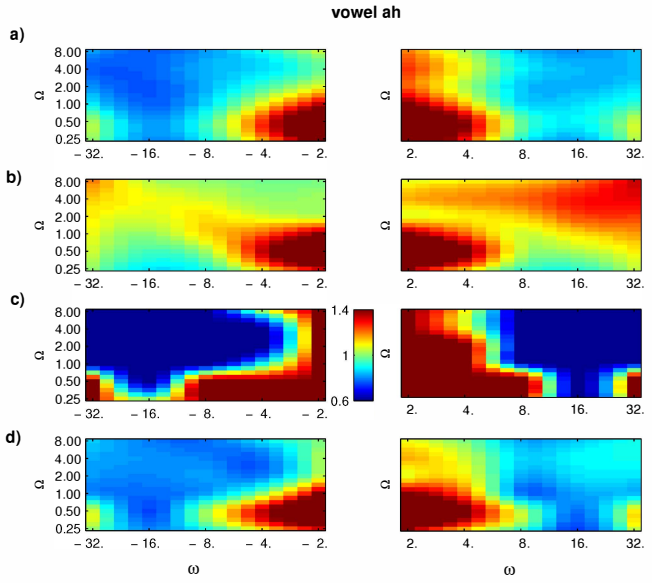

Fig. 3. Gain adaptation for vowel *ah*: a) The plot shows the rate scale representation of target vowel *ah* in clean conditions. b) The plot shows the rate scale representation of the target in white noise at an SNR of 0dB. c) The plot shows the retuned gain values of filters at different rates and scales. The gain values range is [0.6 1.4] as indicated by the bar plot. d) The plot shows the rate-scale plot of the vowel in white noise estimated using the retuned filters.

B. Bandwidth adaptation

Parameters $\sigma_{t_{\omega\Omega}}$ and $\sigma_{f_{\omega\Omega}}$ denote the bandwidths of the gaussians that are used to modulate sinusoidal plane as defined in Equation 3. Both set of parameters are initialized as shown in Equations 9 and 10 accounting for approximately 2 cycles of the sinusoids along time and frequency axis. The patch size used to perform the 2 dimensional convolution is a fixed factor of the bandwidth. Figure 5a shows the a Gabor Filter at a rate of 4Hz and scale of 2 cycles/octave with bandwidth at the initialized values. On increasing the bandwidth, the gaussian

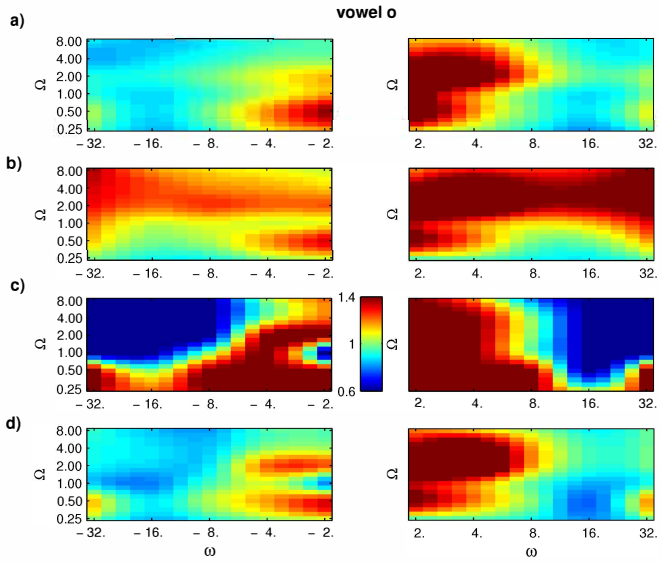

Fig. 4. Gain adaptation for vowel *o*: a) The plot shows the rate scale representation of target vowel *o* in clean conditions. b) The plot shows the rate scale representation of the target in white noise at an SNR of -5dB. c) The plot shows the retuned gain values of filters at different rates and scales. The gain values range is [0.6 1.4] as indicated by the bar plot. d) The plot shows the rate-scale plot of the vowel in white noise estimated using the retuned filters

envelope broadens with energy distributed along more cycles of the sinusoids. Figure 5b shows this effect along the time axis on increasing the bandwidth to $\hat{\sigma}_t = \frac{1}{1.5\omega}$. The opposite effect can be seen in 5c where bandwidth has been retuned to $\hat{\sigma}_t = \frac{1}{2.5\omega}$.

$$\sigma_t = \frac{1}{2\omega} \tag{9}$$

$$\sigma_f = \frac{1}{2\Omega} \tag{10}$$

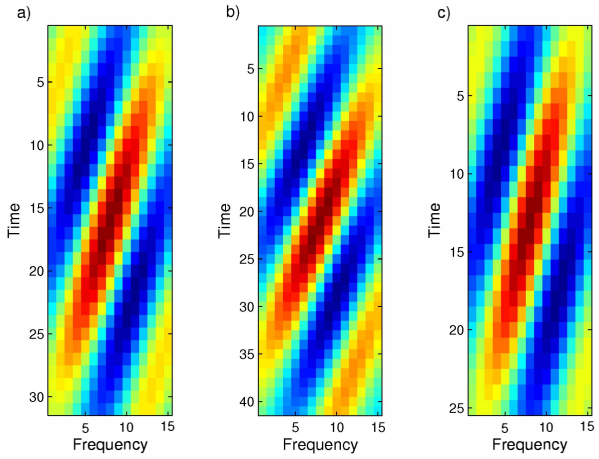

Fig. 5. Gabor filter patch tuned to 4Hz and 2 cycles/octave tuned to different bandwidth. a) $\sigma_t = \frac{1}{2\omega}$, b) $\sigma_t = \frac{1}{1.5\omega}$, c) $\sigma_t = \frac{1}{2.5\omega}$

Figure 6 illustrates the usefulness of adapting bandwidth parameters in a task driven setting. Figure 6a shows the rate-scale representation of the diphthong *au* in clean conditions

and Figure 6b shows the rate-scale representation on adding babble noise at 0dB SNR. It can be seen that on adding babble noise, rate-scale representation of the target signal gets corrupted, introducing energies at high positive rates as well energy at 4Hz on the negative rate axis masking the energies of the diphthong.

Figure 6c shows the proposed change in values of $\sigma_{t_{\omega\Omega}}$, that is bandwidth of the gaussians along the time axis. The parameter value range is $[1/2.5\omega \ 1/1.5\omega]$. The denominator multiplier is indicated by the color bar. It can be seen in Figure 6e that sharpening of filters at rates and scale with strong target presence accentuates the energy of the gaussian around the peak of the sinusoid emphasizing the target. On the same note, broadening the gaussian at rate and scale with strong noise presence dissipates the energy along larger number of cycles of the sinusoid. With lack of clear periodicity in noise, the noise regions in the rate-scale plot get suppressed. These regions are indicated by the magenta ellipse in Figure 6e.

Figure 6d shows the proposed change in values of $\sigma_{f_{\omega\Omega}}$, the bandwidth of the gaussians along the frequency axis. In this case the parameter value range is $[1/2.5\Omega \ 1/1.5\Omega]$ as indicated by the colorbar. The energies of the diphthong along the positive rate axis is mainly due to the narrowband energies of the harmonics in the auditory spectrum. By sharpening the filters along the frequency axis, the target representation in these areas can be accentuated. Along the negative axis though, broadening the filters at rates and scales with strong presence of both target and noise was seen to be beneficial. While in the case of noise, broadening of filters dissipates the energy at these filters, in the case of the target, broadening the gaussian seems capture the broader formants better, resulting in the accentuation of the target. This region is indicated by a black ellipse in Figure 6e. A cosine similarity measure of 0.8226 was estimated between the RSF representation of the clean and the noisy diphthong *ay* on using the default bank of Gabor filters. On retuning the filter bandwidths, the cosine similarity between the clean RSF representation and the new RSF representation improved to 0.8863.

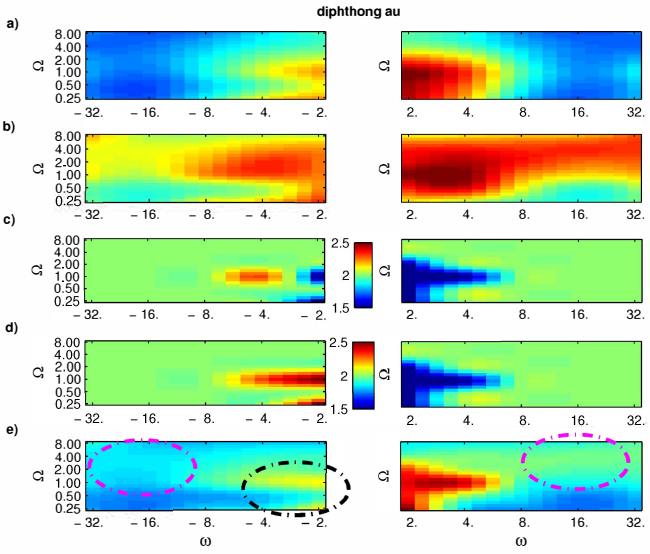

Fig. 6. Bandwidth adaptation for diphthong *au*: a) The plot shows the rate scale representation of target diphthong *au* in clean conditions. b) The plot shows the rate scale representation of the target in babble noise at an SNR of 0dB. c,d) The plot shows the retuned $\hat{\sigma}_{t_{\omega\Omega}}$ and $\hat{\sigma}_{f_{\omega\Omega}}$ values of filters at different rates and scales. The bandwidth values range is $[1/2.5\omega \ 1/1.5\omega]$ and $[1/2.5\Omega \ 1/1.5\Omega]$. The bar plots indicate the denominator multiple. e) The plot shows the rate-scale plot of the *au* in babble noise estimated using the retuned filters. The black ellipse highlights the enhanced target regions and the magenta ellipse indicates the noise suppressed regions.

C. Orientation adaptation

The parameter $\theta_{\omega\Omega}$ offers another layer of flexibility in reshaping the Gabor filters. Figure 7a shows a Gabor filter tuned to 4Hz and 2 cycles/octave at $\theta_{4,2} = 0$ Figure 7b and 7c show the same filter rotated by +10 and -10 degrees by setting $\hat{\theta}_{4,2}$ to $\frac{\pi}{18}$ and $\frac{-\pi}{18}$ respectively.

Figure 8 illustrates the benefit of tuning the orientation for the task of diphthong *ay* in the presence of babble noise at 0dB SNR. Figure 8a shows the rate-scale representation of diphthong *ay* in clean condition, while figure 8b shows the rate scale representation of the speech token in babble noise at 0dB SNR. In both these cases, $\theta_{\omega\Omega} = 0$ for all filters. In clean conditions, as can be seen in figure 8a, energy at low rates and scales dominate the rate-scale representation. Due to the gliding of the formants, the diphthong is also characterized by energies at higher scale values at rates 2Hz and 4Hz. These definitive characteristics are masked on adding noise (figure 8b). In presence of such directionally in the spectrum, one can attempt to re-orient the filters along the direction of the target to enhance the target representation. In the instance of strong directionality exhibited by noise, filters can be retuned to orient away from the noise, thus inhibiting the filter response. Figure 8c shows the proposed retuning of $\theta_{\omega\Omega}$ by at most 8 degrees in either direction. Figure 8d shows the rate-scale representation of the noisy signal on using retuned filters. It can be seen that the signature regions of the diphthong get emphasized, as indicated by the black ellipse. While regions indicated using the magenta ellipse show considerable suppression of noise representation. Further, on estimating the cosine similarity, an increase in cosine similarity from 0.8313 to 0.8714 was observed.

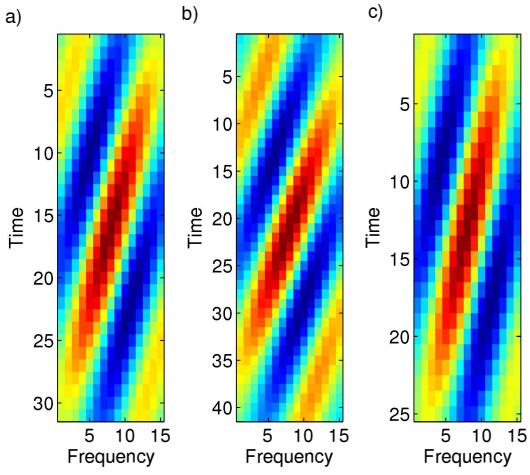

Fig. 7. Gabor filter patch tuned to 4Hz and 2 cycles/octave at different orientation. a) $\theta = 0$, b) $\theta = -\pi/18$, c) $\theta = \pi/18$

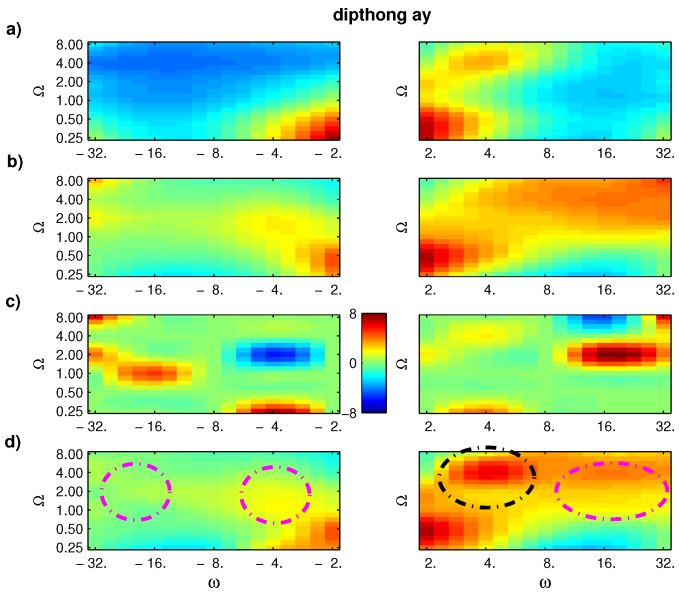

Fig. 8. Theta adaptation for diphthong *ay*: a) The plot shows the rate scale representation of target diphthong *ay* in clean conditions. b) The plot shows the rate scale representation of the target in babble noise at an SNR of 0dB. c) The plot shows the retuned $\hat{\theta}_{t_{\omega\Omega}}$ values. The retuned orientation range is [-8 8] degrees as indicated by the bar plot. d) The plot shows the rate-scale plot of the *ay* in babble noise estimated using the retuned filters. The black ellipse highlights the enhanced target regions and the magenta ellipse indicates the noise suppressed regions.

V. Conclusion

In this work we examined the usefulness of incorporating plasticity for detecting speech tokens in the presence of background noise. In particular, we made a case for employing parameterized Gabor filters to model plasticity. Using specific examples and modulation space representations, we have illustrated the flexibility offered by the parameter tuning approach. An improvement in objective similarity measure was also observed on using the proposed method. While in this work, the parameters have been retuned individually for specific examples, the technique can be extended to scenarios where an optimal set of parameters from Gabor filter parameter space are obtained automatically to enhance target representation/detection. Such a process can be employed to address broader problems like speech activity detection, scene analysis and speech recognition.

References

[1] T. D. Griffiths and J. D. Warren,*What is an auditory object?* Nature Reviews Neuroscience 5(11), pp. 887-892, 2004

[2] L. Miller, M. Escabi, H. Read, and C. Schreiner, *Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex*, J Neurophysiol, vol. 87, no. 1, pp. 516-527, Jan 2002.

[3] I. Nelken, *Processing of complex stimuli and natural scenes in the auditory cortex*, Current opinion in Neurobiology, 14(4), pp 474-490, 2004

[4] I. Nelken and O. Bar-Yosef, *Neurons and Objects: The Case of Auditory Cortex*, Frontiers in Neuroscience 2(1), pp 107-113, 2008

[5] M. Elhilali, S. Shamma, J. Z. Simon and J. B. Fritz, *A Linear System's view of the concept of STRFs*, Handbook of modern techniques in auditory cortex, Chapter 2, 2013

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

[7] T. Ezzat, J. Bouvrie, and T. Poggio, *Spectro temporal analysis of speech using 2d gabor filters*, in INTERSPEECH-2007, pp. 506-509, 2007

[8] J. Fritz, M. Elhilali and S. Shamma, *Adaptive changes in cortical receptive fields induced by attention to complex sounds*, Journal of Neurophysiology, 98(4), pp 2337-2346, 2007

[9] S. Atiani, M. Elhilai, S. David, J. Fritz and S. Shamma, *Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields*, Neuron, 61(3), pp 467-480, 2009

[10] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, *Auditory attention: focusing the searchlight on sound*, vol. 17, no. 4, pp. 437-455, 2007

[11] J. Fritz, M. Elhilali and S. Shamma, *Active listening: Task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex*, Hearing Research, 159-176, 2005

[12] T. Chi, P. Ru and S. Shamma, *Multiresolution spectrotemporal analysis of complex sounds*, Journal of Acoustic Society of America, vol. 118, no. 2, pp 887-906, 2005.

[13] P. Boersma and V. Heuven, *Speak and unSpeak with PRAAT*. Glot International, pp 341-345, 2001

[14] A. Varga, H. J. M. Steeneken, M. Tomlinson and D. Jones, *The NOISEX-92 study on the effect of additive noise on automatic speech recognition*, Speech communication, Volume 12 Issue 3, Jpp 247-251, 1993

[15] K. Patil, M. Elhilali, *Task-driven attentional mechanisms for auditory scene recognition*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp 828-832, 2013

[16] N. Mesgarani and S. Shamma, *Denoising in the domain of spectro-temporal modulations*, EURASIP Journal on Audio, Speech, and Music Processing, Article ID 42357, vol. 2007