SENSORY MAPPING ADAPTATION UNDER MULTIPLE TASK SCENARIOS

Ashwin Bellur, Mounya Elhilali

Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

Demands on auditory perception change constantly with natural changes in everyday acoustic environments. Mechanisms, such as attentional feedback, direct the brain to adapt processing of the incoming signal to maximize its ability to detect the presence of a sound of interest or enhance its representation. These top-down feedback processes induce adaptation of the spectrotemporal representation of incoming sounds in a manner that enhances our ability to perform the desired task. In this work, we propose a computational model to implement and study sensory mapping adaptation under *different* task-demands. We propose a common processing framework to examine how sensory mapping adaptation manifests under different task-driven conditions like speech enhancement and robust speech activity detection. Objective measures of speech enhancement and discrimination are used to quantify the impact of the adaptation under different contexts and its impact on performance outcomes.

Index Terms— Auditory Attention, Adaptation, Spectrotemporal Filters, Speech in Noise, Genetic Algorithm

1. INTRODUCTION

We live in a rich and complex acoustic world, with multiple sources of sound active at every instant of time. Humans are extremely adept at interacting and performing auditory tasks in such a complex acoustic environment. Neurophysiological studies have shed light on some of the processes of the auditory pathway that render the human auditory system so effective [1–3]. Studies have shown that the low dimensional time domain waveform undergoes a series of transformations to obtain a high dimensional representation; wherein the frequency content and the spectrotemporal modulations of the stimulus are encoded [4]. Furthermore, studies show that when performing an auditory task, attentional mechanisms further complement the sensory mapping process. Through use of top down feedback, the sensory mapping process is adapted in a manner that enhances the ability of the auditory system in performing the required task [5–7]. Numerous studies have leveraged the high dimensional sensory mapping processes for feature extraction in audio processing applications [8–10]. Deep belief and convolutional networks inspired by biology [11, 12] have led to remarkable improvement in the performance of data driven speech processing systems.

The focus of this work however is towards the the complementary task-driven top-down attentional mechanisms. There has been a recent body of work that explores different frameworks to model and leverage these attentional mechanisms [13–17]. These attentional models operate on a common underlying principle; an adaptable bio-mimetic sensory mapping or feature extraction process, able to adapt its processing characteristics or tuning properties in a manner that enhances the performance of the task at hand. The driving source behind this adaptation is feedback guided by attention. Most approaches that examine the role of this feedback rely on the same basic principle, but differ in their approach depending on the goal of the system. For instance, the model in Mesgarani et al. focuses on discrimination between arbitrary simple sounds [14] , while work by Carlin, Bellur and colleagues is centered around robust speech activity detection [16, 17]. Patil and Elhilali take a complementary approach to detect auditory scenes [15] while Kalinli and colleagues addresses prominent syllable detection [13]. In this diverse literature, the nature of adaptation varies across frameworks, and spans the continuum from linear optimization [14–16] to nonolinear transformations [17]. The lack of common principles and constraints on these diverse systems makes it challenging to compare manifestations of top-down attentional mechanisms in terms of sensory mapping adaptation, across different tasks.

In this work, we seek to study and compare the outcomes sensory mapping adaptation under different task-driven settings. Hence we expand the framework developed in [17] to 3 different tasks; speech enhancement, speech detection and discriminating between speech and nonspeech under noisy conditions. We develop task relevant feedback to drive the adaptation of the sensory mapping process for each of these tasks, within a single framework. We illustrate the outcomes of adaptation under different task-driven scenarios and show that the spectrotemporal modulation space adapts in distinct interesting ways in order to enhance the performance of the corresponding task. We also compare the performance of these task-driven systems in achieving speech enhancement and speech activity detection.

This work was supported by the National Institutes of Health under Grants R01HL133043, ONR N000141010278, N000141612045 and N000141210740.

2. SENSORY MAPPING

The sensory mapping process is divided into 2 stages. In stage one, the sound signal is transformed into a time-frequency representation by passing it through a model of the auditory periphery as developed in [4]. We will refer to the time-frequency representation as the auditory spectrogram and notate it as $y(t, f)$. Stage two is an adaptable sensory mapping process based on the processes observed in the cortical regions of the mammalian auditory pathway. A filter bank of parameterized 2-dimensional Gabor filters is used to model the spectrotemporal receptive fields of auditory neurons in the cortical regions [18]. The bank of filters, notated as $g = \{g_1, \ldots, g_m\}$, spans the spectrotemporal modulation space with individual g_k defined as shown in equation 1.

$$g_k(t, f) = \frac{\alpha_k}{2\pi\sigma_{t_k}\sigma_{f_k}} e^{-\frac{1}{2}\left(\frac{t_1^2}{\sigma_{t_k}^2} + \frac{f_1^2}{\sigma_{f_k}^2}\right)} e^{2\pi j(\omega_k t + \Omega_k f)} \quad (1)$$

where $t_1 = tcos(\theta_k) + fsin(\theta_k)$ and $f_1 = -tsin(\theta_k) + fcos(\theta_k)$. σ_{t_k} and σ_{f_k} denote the bandwidths of the Gaussians of the k^{th} Gabor filter along time and frequency direction respectively. θ_k represents 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.

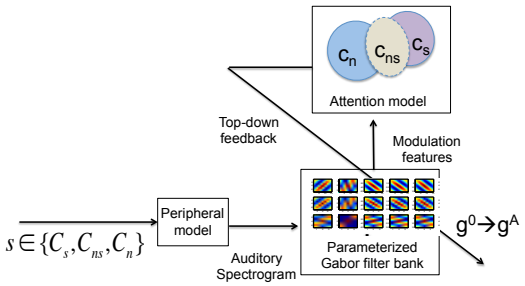

Fig. 1. Task-driven adaptation framework

The bank of filters is convolved with the auditory spectrogram over both time and frequency to encode the frequency and spectrotemporal modulation information in the spectrogram. The response is the collapsed over time to obtain the rate-scale-frequency (RSF) representation of the signal; with each of the individual Gabor filter g_k encoding a particular rate ω_k and scale Ω_k. Temporal information is still retained as slow temporal modulations encoded by filters of varying rates. The response to each filter is obtained as shown in equation 2. The ensemble response to the bank of filters is defined as $E = [T_1, T_2; \ldots, T_m] \in \mathbb{R}^{m*n_f}$, a stacked representation of the response to the m filters, where channels n_f refers to the number of frequency channels in the auditory spectrogram.

$$T_k(f) = \int |y(t, f) *_{tf} g_k(t, f)| dt \quad (2)$$

3. TASK-DRIVEN ADAPTATION

Sensory mapping adaptation in this framework is achieved by retuning the Gabor parameters. Given a bank of filters with the default set of parameters, notated as $g^0 = \{g_1^0, g_2^0, \ldots, g_m^0\}$, the goal of the top-down feedback process is to estimate a set of retuned filters $g^A = \{g_1^A, g_2^A, \ldots, g_m^A\}$, as determined by the task at hand. The framework is as shown in figure 1. In order to perform the adaptation (g^0 to g^A), we use the genetic algorithm as proposed in [17]. Genetic algorithm presents an elegant way to search the Gabor filter bank parameter space for the optimal set of parameters. The algorithm is initialized with the default parameter set as a *member* of the first *generation*. The algorithm then propagates through multiple generations, with each generation having fitter members than the previous generation; members in this context being parameter sets (g) within a prescribed range. The fittest member of the final generation is the desired set of filters g^A. The manner in which the algorithm propagates from generation to generation is as detailed in [17]. The fitness measure is key here and is defined on the basis of the task being performed.

3.1. Speech enhancement

Given stimuli from the clean speech class (C_s), ensemble responses with default filters are estimated. Next, given noisy speech stimuli, distorted versions of the clean speech stimuli, we seek to enhance the speech representation in the RSF space. In order to achieve this, the fitness measure is defined as shown in equation 3. Fitness measure f_{ENH} is the mean euclidean distance between the clean speech response and the corresponding noisy speech representation in the RSF space, using the default and the adapted filters respectively. $E_{s_j}^0$ represents the RSF representation for the j^{th}clean speech stimulus estimated using the original filters. $E_{ns_j}^A$ represents the RSF representation for the corresponding noisy speech stimulus obtained using the adapted filters. It was shown in [19], that such a metric in the RSF closely matches error rates of human listeners in various noisy conditions.

$$f_{ENH} = \frac{1}{J}\sum_{j=1}^{J}(E_{s_j}^0 - E_{ns_j}^A)^2 \quad (3)$$

3.2. Speech detection

While in the previous case, we used the clean speech stimulus as a template to enhance speech in noisy conditions, for the detection task we use a statistical representation, a Gaussian mixture model (GMM) to represent clean speech in the spectrotemporal modulation space. Given a set of clean speech stimulus, the rate-scale-frequency response is first estimated (E). Then the tensor singular value decomposition (TSVD) is used to reduce the number of dimensions of the RSF representation while ensuring that certain percentage of the variance is retained [20]. Gaussian mixture model is then estimated using this reduced-dimensioned representation (notated as V).

The task in this case is to detect presence of speech even in noisy conditions. This is formulated using the fitness measure defined in equation 4. The purpose of using such a fitness measure is to obtain a retuned set of filters g^A that maximizes the average likelihood of the noisy speech samples with respect to the clean model. M_s^0 represents the GMM estimated using clean speech with the default filters g^0. $P(V_{ns_j}^A|M_s^0)$ is the likelihood value of the adapted representation of the noisy speech stimulus with respect to the clean speech GMM M_s^0.

$$f_{DET} = \frac{1}{J}\sum_{j=1}^{J} P(V_{ns_j}^A|M_s^0) \tag{4}$$

3.3. Speech and nonspeech discrimination

Under this task setting, we are seeking to enhance discriminability between low SNR speech and nonspeech classes through sensory mapping adaptation. First we estimate GMMs for clean speech and nonspeech classes as described above. Next, we define the loglikelihood ratio (LLR) as shown in equation 5. V_c^A is the feature extracted from a stimulus belonging to either noisy speech or nonspeech class using the adapted filters. M_s^0 and M_n^0 denote clean speech and nonspeech GMMs estimated using the default sensory mapping procedure.

$$LLR = log\left(\frac{P(V_c^A|M_s^0)}{P(V_c^A|M_n^0)}\right) \tag{5}$$

The goal in this task is to adapt the sensory mapping process in a manner that allows the clean speech and nonspeech GMMs to discriminate between low SNR speech and nonspeech, even in mismatched conditions. In order to do so, d-prime as defined in equation 6 is used as the fitness measure.

$$f_{DIS} = \frac{\mu_{ns} - \mu_n}{\sqrt{(\frac{1}{2}(\sigma_{ns}^2 + \sigma_n^2))}} \tag{6}$$

Symbols μ_c and σ_c denote the mean and standard deviation respectively of the LLR values. $c = ns$ denotes noisy speech samples and $c = n$ denotes samples from the nonspeech class.

4. EXPERIMENTS AND RESULTS

The outcomes of the adaptation under task-driven settings were studied using a variety of clean speech, noisy speech and nonspeech classes. Data from the TIMIT database [21] was used as clean speech data. Cafe noise from QUT-Noise database [22], and sounds belonging to *emergency* class from the BBC sound effects database [23] were used as the nonspeech classes and as sources of additive noise. Along with additive noise, 2 nonlinear distortions of speech, reverberated speech and speech with phase jitter [19] were also used to study task-driven adaptation. The genetic algorithm was run separately for each of the noise cases. For example, in order to study the cafe-noise scenario, noisy speech data for adaptation is created using cafe-noise as additive noise and cafe-noise GMM model is used as the nonspeech model for the discrimination task. For the nonlinear distortions, random sampling of sounds from the BBC sound effects database were used to create the nonspeech GMMs. A separate held out dataset from these databases were used in all cases to study and test the proposed systems.

Gabor filter bank g^0 were estimated at rates ranging from 2 Hz to 32 Hz and scales ranging from 0.25 to 8 cycles/octave. The default parameters were initialized as follows $\forall \omega, \Omega$:
$\sigma_{t_{\omega\Omega}} = \frac{1}{2\omega}$, $\sigma_{f_{\omega\Omega}} = \frac{1}{2\Omega}$, $\theta_{\omega\Omega} = 0$ and $\alpha_{\omega\Omega} = 1$

A range of values in vicinity of the default values of the parameters were used as the parameter space. $\boldsymbol{\sigma_t} = [\frac{1}{1.5\omega}\ \frac{1}{2.5\omega}]$, $\boldsymbol{\sigma_f} = [\frac{1}{1.5\Omega}\ \frac{1}{2.5\Omega}]$, $\boldsymbol{\theta} = [-3\ \ 3]$ (in degrees), $\boldsymbol{\alpha} = [0.5\ \ 1.5]$. The genetic algorithm then operates within the limited search space to determine g^A under the different task-driven settings. We obtain a task specific adapted sensory mapping process using the prescribed fitness measures for each of the different tasks as described in section 3.

4.1. Cosine similarity

Figure 2 shows the average cosine similarity between the clean speech RSF representation and the noisy speech RSF representation before (g^0) and after adaptation (g^A). As can be seen in figure 2, the adapted representation g_{ENH}^A, estimated under the enhancement setting, performs best and is closest to the clean speech RSF representation across all noise cases, with marked improvement under low SNR conditions. The g_{DET}^A filters though, obtained under the speech detection framework, performs well in low SNR conditions for cafe-noise (babble like noise) and reverberation conditions, with deterioration in performance for the emergency noise class and phase jitter. g_{DIS}^A performs similar to the default sensory mapping process (g^0) except for the reverberation case. It is evident that the adaptation manifests very differently under different task-driven conditions. Enhanced ability in detecting speech or discriminating between speech and noise, does not necessarily lead to improved enhancement of speech.

4.2. Equal error rate

Table 1 shows the equal error rates (EER) estimated using noisy speech and nonspeech stimuli under different task-driven settings. The EERs were obtained using the LLR values estimated as defined in equation 5 for a held out set of noisy speech and nonspeech data using the GMM models. As can be seen, the discriminatory filters g_{DIS}^A perform best across all 4 noise conditions, with considerable improvement over the default setup g^0. The adapted processes obtained under the enhancement task g_{ENH}^A and the detection task g_{DET}^A are not consistent in their performance across different noise conditions.

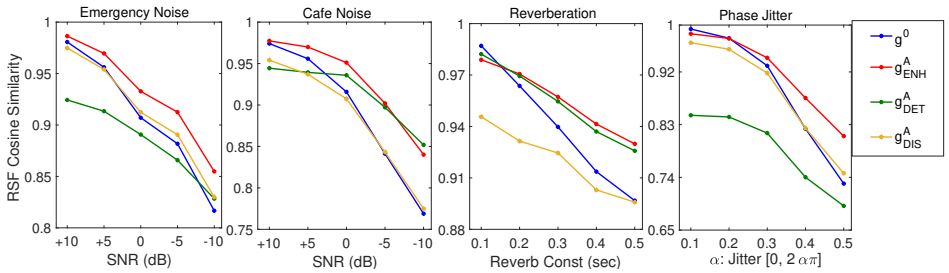

Fig. 2. Average cosine similarity between clean speech and the corresponding noisy counterpart.

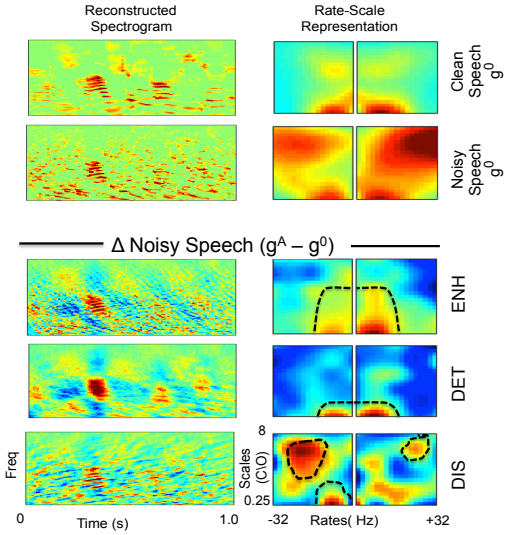

Fig. 3. Reconstructed auditory spectrograms on the left and rate-scale representation on the right. Row 1 and 2 show the clean speech and noisy speech representations respectively, using g^0. Row 3, 4 and 5 show the difference between the adapted and original noisy speech representations.

Table 1. Equal Error Rate

Noise condition	g^0	g^A_{ENH}	g^A_{DET}	g^A_{DIS}
Emergency	14.74	12.35	18.33	**10.05**
Cafe	26.05	25.95	22.67	**21.12**
Reverb	9.40	4.60	4.55	**2.40**
Jitter	28.81	30.05	31.23	**14.41**

4.3. Rate-scale analysis

In order to understand the results obtained, we analyze the reconstructed auditory spectrograms (using the Gabor filter bank and its responses) and the corresponding rate scale representations under different task-driven settings in figure 3. The rate-scale (RS) representations are obtained by averaging the RSF representation in equation 2 over frequency and reshaping them such that the x-axis denotes rates and y-axis scales. Row 1 is the reconstructed clean speech spectrogram using g^0 and the corresponding rate-scale energy spread. Row 2 is the noisy speech representation using g^0 with additive noise from the cafe-noise class. Rows 3 to 5 show the difference between reconstructed spectrograms and the rate-scale energy spread, on using task specific filters g^A and the default filters g^0. Red areas indicate enhancement and blue areas indicate suppression. Under the speech enhancement setting (g^A_{ENH}), it can be seen in the RS space that the spectrotemporal modulations pertaining to speech are emphasized (regions within the dotted black lines in row 3). While this leads to better similarity measures, this implies areas where speech and nonspeech overlap are also retained, hence impeding its ability to discriminate between speech and nonspeech. Under the detection setting (g^A_{DET}), adaptation leads to a sparser representation of the auditory spectrogram with focus on few key speech modulations as indicated by the dotted black lines in row 4. While this sharp focus improves the ability to detect speech, it does not necessarily result in enhanced perception of speech in all conditions. Hence the drop in the similarity measures under high SNR conditions where sparsity results in poorer representation of speech in the RSF space. In the discrimination case (g^A_{DIS}, row 5), dotted black regions in the RS representation highlight the non-overlapping speech and nonspeech regions that are emphasized. While this leads to improved EERs, it does not lead to consistent improvement in similarity measures, as even the distinct nonspeech regions are retained, while the overlapping regions are suppressed. Only in cases with speech like noise conditions, does it lead to improvement in similarity measures.

5. CONCLUSION

In this work, within one single framework, we studied task-specific sensory mapping adaptation for representation of speech in noisy settings. Using a feedback driven nonlinear adaptation framework we showed that depending on the task, very distinct and specific regions of the spectrotemporal modulation space is adapted. In the speech enhancement task the focus was on preserving and emphasizing speech regions. Whereas in the detection and discrimination task, very specific sparse regions of the spectrotemporal modulation space was enhanced while the rest was suppressed. Characteristic nature of the adaptation under different task-driven conditions was further illustrated by estimating objective measures of enhancement (similarity to clean speech) and speech discrimination (equal error rates).

References

[1] Jos J Eggermont, "Between sound and perception: reviewing the search for a neural code," *Hearing research*, vol. 157, no. 1-2, pp. 1–42, 2001.

[2] Arthur N. Popper and Richard R. Fay, Eds., *The Mammalian Auditory Pathway: Neurophysiology*, vol. 2 of *Springer Handbook of Auditory Research*, Springer New York, New York, NY, 1992.

[3] K T Hill and L M Miller, "Auditory Attentional Control and Selection during Cocktail Party Listening," *Cerebral cortex (New York, N.Y.: 1991)*, vol. 20, no. 3, pp. 583–590, mar 2009.

[4] T Chi, P Ru, and S A Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[5] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Adaptive Changes in Cortical Receptive Fields Induced by Attention to Complex Sounds," *Journal of Neurophysiology*, vol. 98, no. 4, pp. 2337–2346, 2007.

[6] P Yin, J B Fritz, and S A Shamma, "Rapid spectrotemporal plasticity in primary auditory cortex during behavior," *The Journal of neuroscience*, vol. 34, no. 12, pp. 4396–4408, mar 2014.

[7] S Shamma and J Fritz, "Adaptive auditory computations," *Current opinion in neurobiology*, vol. 25, pp. 164–168, apr 2014.

[8] Martin Heckmann, Xavier Domont, Frank Joublin, and Christian Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Communication*, vol. 53, no. 5, pp. 736–752, 2011.

[9] Michael Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[10] Nima Mesgarani, Malcolm Slaney, and Shihab Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 920–930, 2006.

[11] Shuo-Yiin Chang and Nelson Morgan, "Robust CNN-based speech recognition with Gabor filter kernels.," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] Jui-Ting Huang, Jinyu Li, and Yifan Gong, "An analysis of convolutional neural networks for speech recognition," in *ICASSP*, 2015, pp. 4989–4993.

[13] Ozlem Kalinli and Shrikanth Narayanan, "A top-down auditory attention model for learning task dependent influences on prominence detection in speech," in *International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 3981–3984.

[14] N Mesgarani, J Fritz, and S Shamma, "A computational model of rapid task-related plasticity of auditory cortical receptive fields," *Journal of computational neuroscience*, vol. 28, no. 1, pp. 19–27, feb 2010.

[15] Kailash Patil and Mounya Elhilali, "Task-driven attentional mechanisms for auditory scene recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. may 2013, pp. 828–832, IEEE.

[16] Michael A. Carlin and Mounya Elhilali, "Modeling attention-driven plasticity in auditory cortical receptive fields," *Frontiers in computational neuroscience*, vol. 9, pp. 106, 2015.

[17] Ashwin Bellur and Mounya Elhilali, "Feedback-Driven Sensory Mapping Adaptation for Robust Speech Activity Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 481–492, mar 2017.

[18] Tony Ezzat, Jake V Bouvrie, and Tomaso Poggio, "Spectro-temporal analysis of speech using 2-d Gabor filters.," in *Interspeech*, 2007, pp. 506–509.

[19] Mounya Elhilali, Taishih Chi, and Shihab A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, no. 2-3, pp. 331–348, 2003.

[20] L De Lathauwer, B De Moor, J Vandewalle, L De Lathauwer, B De Moor, and J Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, pp. 1253–1278, 2000.

[21] J S Garofolo, L F Lamel, W M Fisher, J G Fiscus, D S Pallett, N L Dahlgren, and V Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, 1993.

[22] David B Dean, Sridha Sridharan, Robert J Vogt, and Michael W Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech*, 2010.

[23] Http://www.sound-ideas.com/bbc.html, "The BBC Sound Effects Library," 1990.