Feedback-Driven Sensory Mapping Adaptation for Robust Speech Activity Detection

Ashwin Bellur and Mounya Elhilali, *Senior Member, IEEE*

***Abstract*—Parsing natural acoustic scenes using computational methodologies poses many challenges. Given the rich and complex nature of the acoustic environment, data mismatch between train and test conditions is a major hurdle in data-driven audio processing systems. In contrast, the brain exhibits a remarkable ability at segmenting acoustic scenes with relative ease. When tackling challenging listening conditions that are often faced in everyday life, the biological system relies on a number of principles that allow it to effortlessly parse its rich soundscape. In the current study, we leverage a key principle employed by the auditory system: its ability to adapt the neural representation of its sensory input in a high-dimensional space. We propose a framework that mimics this process in a computational model for robust speech activity detection. The system employs a 2-D Gabor filter bank whose parameters are retuned offline to improve the separability between the feature representation of speech and nonspeech sounds. This retuning process, driven by feedback from statistical models of speech and nonspeech classes, attempts to minimize the misclassification risk of mismatched data, with respect to the original statistical models. We hypothesize that this risk minimization procedure results in an emphasis of unique speech and nonspeech modulations in the high-dimensional space. We show that such an adapted system is indeed robust to other novel conditions, with a marked reduction in equal error rates for a variety of databases with additive and convolutive noise distortions. We discuss the lessons learned from biology with regard to adapting to an ever-changing acoustic environment and the impact on building truly intelligent audio processing systems.**

***Index Terms*—Adaptation, gabor filters, genetic algorithm, spectrotemporal filters, speech activity detection.**

I. Introduction

THE acoustic world we inhabit is a rich one, often composed of multiple sound sources. Developing computational techniques to parse a complex acoustic scene poses numerous challenges. Given the increasing desire to process such real world data for tasks like speech detection and recognition, data tagging, source separation and coding, there is great deal of emphasis on developing robust computational methodologies capable of working with complex real-world acoustic signals.

One of the main issues when dealing with real-world acoustic signals using data-driven computational techniques, is the problem of data mismatch. That is, a mismatch between the statistics captured by the system during the training phase and statistics of data used during testing. This problem often arises owing to the rich nature of the acoustic environment and the inability to include all possible scenarios during the training phase. While data-driven state of the art systems, be it in speech detection or recognition or scene analysis, are remarkably accurate in matched conditions, the performance drops rapidly under mismatched conditions. In contrast, human listeners are amazingly adept at dealing with such complex acoustic scenes, especially in adapting to changing and novel acoustic environments. A commonly cited example is that of a cocktail party, where we are able to communicate with notable ease in the midst of music, clinking of glasses and loud background chatter [1]. Studies of the neurophysiology of the mammalian auditory system have shed light on some of the processes that render the auditory system efficient in complex soundscapes [2]–[4]. The goal of this study is to leverage some of these processes for building a robust data-driven audio processing system.

Our knowledge of brain processes reveals that the time-domain sound signal, a low-dimensional vector, undergoes a series of transformations along various stages of the auditory system, akin to a mapping onto high-dimensional space [5]. This mapping captures modulations or variations in the signal along both time and frequency. By projecting the signal onto this high-dimensional spectro-temporal modulation space, different components of the acoustic scene are etched out, effectively occupying different sub-regions of the space; which in turn enable the brain to effectively parse the acoustic scene [6]. Complimentary to this high-dimensional mapping are adaptation mechanisms that allow the biological system to re-tune its filtering properties in a direction that facilitates the segregation of target sounds from background distractors. Particularly, when listening to a specific sound of interest in a scene, mechanisms of selective-attention provide a feedback control that adapts the sensory mapping to enhance the representation of the target regardless of competing masker sounds [7]–[9]. This feedback-driven adaptation of sensory processing gives the brain a notable advantage over engineering systems, allowing it to adapt to its environment even in novel, previously unseen acoustic surrounds.

There have been numerous efforts to develop computational algorithms to incorporate bio-inspired processes of high-dimensional mapping into audio technologies [10]–[13]. In fact,

Manuscript received May 26, 2016; revised September 29, 2016; accepted November 30, 2016. Date of publication December 13, 2016; date of current version January 16, 2017. This work was supported by the National Institutes of Health under Grants R01HL133043, ONR N000141010278, N000141612045 and N000141210740. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui.

The authors are with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: abellur1@jhu.edu; mounya@jhu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2639322

recent trends using deep belief and convolutional networks effectively model the acoustic space using filters that reflect a similar tiling of the spectrotemporal space akin to the high-dimensional mapping in the auditory system [14], [15]. These bio-inspired and representation learning techniques have been applied for various tasks like speech recognition [11], [16]–[18], speech activity detection [19]–[21], source separation [22], [23], scene recognition [24], [25] and timbre recognition [26], [27].

On the other hand, there have been relatively fewer efforts that have sought to leverage the complimentary task-driven adaptation phenomenon [28], [29]. The main focus of the current work is to develop a framework that not only performs the prerequisite high-dimensional sensory mapping, but also adapts in a task-driven setting so as to enhance the robustness of the system. Given the effectiveness of these biological processes in facilitating speech processing even in the most adverse acoustic environments [30], we develop a system for robust speech activity detection (SAD) that can operate in novel noisy soundscapes. The next section provides an overview of the proposed framework and outlines how the system integrates processes of sensory mapping and adaptation to achieve a robust representation of speech signals. Section III provides details of the implementation of the high-dimensional mapping of the acoustic signal as well as statistical modeling of speech and non-speech sound classes. Section IV then details the proposed methodology to incorporate feedback driven adaptation of the sensory mapping process. In section V, we provide specifics of the proposed SAD system as well as databases employed to validate it. We report the results in Section VI and conclude with a discussion in Section VII.

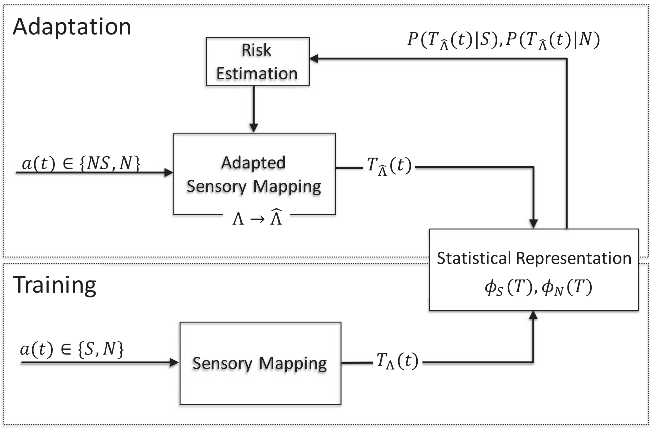


Fig. 1. Training: $a(t)$ is the acoustic signal from clean speech (S) and non-speech (N) classes during the training phase. T_{Λ} denotes the sensory mapping process ϕ denotes the conditional distributions. Adaptation: data $a(t)$ is from noisy speech (NS) and non-speech classes. $T_{\hat{\Lambda}}$ denotes the adapted sensory mapping process. Risk estimated using posterior probabilities $P(T_{\hat{\Lambda}}(t)|S)$ and $P(T_{\hat{\Lambda}}(t)|N)$ with respect to the original models ϕ; goal of iterative adaptation is to minimize risk

II. Overview of the Adaptive Sensory Mapping Framework

The proposed system comprises of three key components: (i) Sensory mapping, implemented using a hierarchy of time-frequency short-term analyses followed by multi-resolution filtering via an array of parameterized two-dimensional Gabor filters. The resultant representation is a high-dimensional feature transformation that encodes spectrotemporal modulations of the incoming signal and serves as an effective biomimetic approximation of the sensory mapping observed in the mammalian system. (ii) Statistical modeling, which learns generative models of speech and non-speech data in the high-dimensional modulation space. (iii) Adaptation, which re-tunes the Gabor filters using feedback from the statistical models. Effectively, this stage allows the system to tackle mismatched conditions by changing the sensory mapping in order to best fit the statistical models of speech and non-speech. Using held-out noisy speech data, the sensory mapping is adjusted while keeping the statistical models fixed, with a goal to maintain discriminability between speech and noise/non-speech even in adverse mismatched conditions.

Fig. 1 shows an overview of the proposed system. Our working hypothesis is as follows: during the training phase, statistical representations of the speech class $\phi(T|S)$ and non-speech class $\phi(T|N)$ are estimated over sensory space T_{Λ}. These representations serve as fixed accounts of the most-informative regions of speech and noise in the modulation space. The subspaces are assumed to be well-separated given the high-dimensional nature of the mapping T_{Λ} as well as the discernible statistical differences between clean speech and noise. However, under low signal-to-noise ratio (SNR) conditions, there is a larger overlap in the regions occupied by noisy speech (NS) and non-speech signals (N). This results in a larger *risk* of misclassification, which can be computed using conditional probabilities $P(T_{\Lambda}(t)|S)$ and $P(T_{\Lambda}(t)|N)$, given labeled data [31]. In order to mitigate this risk, we propose to use a measure of misclassification risk to re-tune our sensory mapping in order to keep the risk minimal and effectively maximize the separability of the two classes in a direction that matches their original statistical models $\phi(T|S)$ and $\phi(T|N)$. The risk measure acts as a feedback control that adaptively changes the sensory mapping indexed by Λ onto a modified space $\hat{\Lambda}$. This process is achieved iteratively with the reduced risk measure estimated at every iteration using $P(T_{\hat{\Lambda}}(t)|S)$ and $P(T_{\hat{\Lambda}}(t)|N)$ with respect to the original models ϕ. The process leverages known physiological projections from higher cognitive brain regions onto sensory cortex which adaptively modulate tuning properties of auditory filters in a direction that maximizes figure/ground segregation and enhances the robust encoding of target sounds despite presence of severe interference [9], [32]. We hypothesize that this risk minimization procedure results in an emphasis of unique speech and non-speech modulations, as represented by the original fixed models (ϕ). Consequently, by tuning the sensory mapping to put a spotlight on unique signatures of speech and non-speech features, the system is able to generalize to unseen conditions and operate robustly under various distortions. As will be shown in this work, this adaptation process results in a nonlinear alteration of the sensory mapping Λ onto the new mapping $\hat{\Lambda}$ that enhances the discriminability between speech and non-speech sound classes.

It is important to note that the proposed framework differs from other model-adaptation methods commonly employed for addressing data mismatch, like maximum a posteriori (MAP) [33] and maximum likelihood linear regression (MLLR) [34]. Unlike these aforementioned techniques, the statistical speech and non-speech models used in the current approach are kept intact while changing the feature space to best match these models. Additionally, the proposed work is different from multi-condition training neural networks where labeled noisy speech and non-speech data are used to train robust neural network systems [20], [35]–[37]. While the basis or weights estimated in these systems are shown to be robust, they are still constrained by the nature of labeled training data and are susceptible to data-mismatch. In our current approach, feedback from statistical models is used to direct the feature space to primarily highlight *known* disparate speech and non-speech regions as represented by the statistical models.

III. Sensory Mapping and Modeling of Speech and Non-Speech Classes

A. *Bio-Mimetic Sensory Mapping*

The incoming sound waveform undergoes a series of transformations along the auditory pathway that extract informative cues about the signal and sound sources present in the environment [3]. These transformations, dubbed bottom-up processes in reference to their feed-forward nature, project the time-domain signal onto a different auditory space that facilitates tasks of sound detection, recognition and segregation [38]–[40].

In the current work, the acoustic waveform is first mapped onto a time-frequency spectrogram using a number transformations mimicking processes in the mammalian auditory periphery. Details of this early transformation can be found in [5] but are summarized next. The incoming signal $a(t)$ is first filtered through an array of asymmetric, constant-Q, band-pass filters $h_c(t, f)$ which span 5.3 octaves on a logarithmic scale, starting at frequency of 180 Hz. The filters $h_c(.)$ are pre-defined to match neurophysiologically-derived cochlear filters. The filtered frequency-dependent signals are then half-wave rectified, integrated over a short time window $w(t; \tau) = e^{-t/\tau} u(t)$ where $\tau = 8ms$, and then compressed using a cubic root compression. Equation (1) succinctly summarizes these steps, where $*$ denotes convolution with respect to time.

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

Next, the resultant time-frequency spectrogram $y(t, f)$ is further analyzed using a filter bank of two-dimensional Gabor filters (the choice of parameterized Gabor filters will become clear later in the text). These filters can be considered as linear approximations of the transfer functions of auditory neurons in the central stages of the mammalian auditory pathway [41], [42]. By having each filter tuned to a particular spectral modulation (or scale) Ω and temporal modulation (or rate) ω, this filtering process highlights the spectro-temporal modulations present in the signal and effectively tiles the spectrotemporal modulation space. The Gabor filters are parameterized as shown in equation (2).

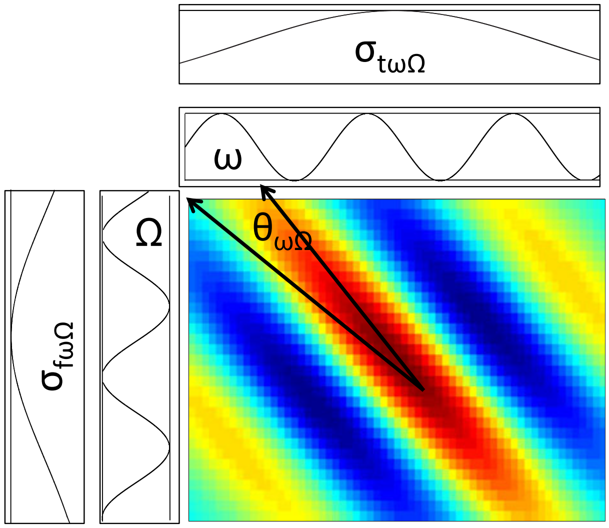


Fig. 2. Two dimensional Gabor filter at rate ω and scale Ω. σ_t is the bandwidth of the Gaussian along the time axis. σ_f is the bandwidth of the Gaussian along the frequency axis. θ denotes the orientation of the main lobe of the Gabor filter

$$F(\omega, \Omega, t, f|\Lambda) = \frac{\alpha_{\omega\Omega}}{2\pi\sigma_{t_{\omega\Omega}}\sigma_{f_{\omega\Omega}}} e^{-\frac{1}{2}\left(\frac{t_1^2}{\sigma^2_{t_{\omega\Omega}}} + \frac{f_1^2}{\sigma^2_{f_{\omega\Omega}}}\right)} e^{2\pi j(\omega t + \Omega f)} \qquad (2)$$

where $t_1 = t\cos(\theta_{\omega\Omega}) + f\sin(\theta_{\omega\Omega})$ and $f_1 = -t\sin(\theta_{\omega\Omega}) + f\cos(\theta_{\omega\Omega})$. The parameters in equation (2) are:

1) ω: in Hz represents the rate (temporal modulations) and Ω: in cycles/octave is the scale (spectral modulations) which determine the variations of the filter in time and frequency. Gabor filters can be downward- or upward-selective with upward-selective filters denoted using negative rate values.
2) $\sigma_{t_{\omega\Omega}}$ and $\sigma_{f_{\omega\Omega}}$ denote the bandwidths of the Gaussians of the Gabor filters along time and frequency direction respectively.
3) $\theta_{\omega\Omega}$ specifies the orientation of the main lobe of the Gabor filter.
4) $\alpha_{\omega\Omega}$ is an additional scalar gain. The gain term is used to suppress or enhance the output of the filter.

The parameters are collectively represented as a vector Λ where $\Lambda = \{\sigma_{t_{\omega\Omega}}, \sigma_{f_{\omega\Omega}}, \theta_{\omega\Omega}, \alpha_{\omega\Omega}\}$. Fig. 2 shows a gabor filter tuned to a rate (ω) of $2Hz$ and scale (Ω) of 1 cycle/octave.

The auditory spectrogram obtained using equation (1) is convolved over both time and frequency (denoted as $\otimes$) with the bank of Gabor filters (Eq. (3)). The output of this stage yields a projection onto the modulation space, of which only the magnitude is preserved.

$$R(\omega, \Omega, t, f|\Lambda) = |Y(t, f) \otimes F(\omega, \Omega, t, f|\Lambda)| \qquad (3)$$

The output $R(.)$ is further collapsed along the time axis to obtain a Rate-Scale-Frequency (RSF) representation $T(f, \omega, \Omega|\Lambda)$ as shown in equation (4). While the time axis is effectively removed from the final feature representation over the duration of the analysis frame, temporal information is still maintained

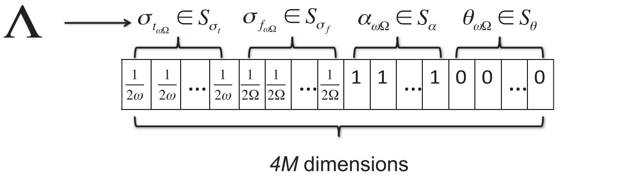

Λ
$\sigma_{t_{\omega\Omega}} \in S_{\sigma_t}$
$\sigma_{f_{\omega\Omega}} \in S_{\sigma_f}$
$\alpha_{\omega\Omega} \in S_{\alpha}$
$\theta_{\omega\Omega} \in S_{\theta}$
$\frac{1}{2\omega}$ $\frac{1}{2\omega}$ ... $\frac{1}{2\omega}$ $\frac{1}{2\Omega}$ $\frac{1}{2\Omega}$ ... $\frac{1}{2\Omega}$ 1 1 ... 1 0 0 ... 0
4M dimensions

Fig. 3. Population of the first generation is initialized as the default 240 dimensional filter bank parameter vector. Each tuple can take any of the five feasible values leading to a search space of $(4M)^C$.

as slow temporal modulations are represented along the rate axis (ω).

$$T(\omega, \Omega, f|\Lambda) = \int R(\omega, \Omega, t, f|\Lambda)dt \tag{4}$$

B. Statistical Models of Speech and Non-Speech Classes

Speech and non-speech classes are fairly separable in the modulation space. In this work, we employ a generative representation using Gaussian mixture models (GMMs) to capture the statistics of both classes. Given data from clean speech or non-speech classes, frame-wise RSF ($T(\omega, \Omega, f|\Lambda)$) features are first extracted. The features extracted are of dimensions [rates × scales × frequency]. In order to facilitate building of GMMs, Tensor Singular Value Decomposition (TSVD) [43] is used to reduce the number of dimensions of the RSF representation while ensuring that certain percentage of the variance is retained. Using the reduced-dimension feature vectors (represented as V_Λ), respective clean speech and non-speech GMMs can be estimated.

IV. Feedback-Driven Adaptation

The GMM models developed earlier result in a degree of separability between speech and non-speech classes as quantified by the Log likelihood ratio (LLR) (Equation (5)), with speech as the null hypothesis. Let $P(V_\Lambda|\Phi)$ denote the posterior probability of the feature V_Λ, extracted from a frame of data with respect to the GMM model Φ. Φ_s and Φ_n, which represent the clean speech and non-speech GMM models, respectively.

$$LLR = \log \frac{P(V_\Lambda|\Phi_s)}{P(V_\Lambda|\Phi_n)} \tag{5}$$

Following the framework presented in Fig. 1, an adaptation stage performs a transformation of the mapping Λ onto a new mapping $\hat{\Lambda}$. The need for adaptation can be motivated by looking at Fig. 4. The dotted histograms in Fig. 4(b), are the LLR histograms of non-speech and speech regions, estimated before adaptation, from a mismatched acoustic signal of 120 seconds duration (Fig. 4(a)). As can be seen from the histograms at different SNR, they are fairly separable when speech is present at a high SNR of 15 dB to 5 dB. However, at lower SNRs (0 dB to -10 dB), the clean speech model is no longer a good representation of the statistics of the speech signal. This leads to a large overlap between LLR values of noisy speech and non-speech, implying a higher *risk* of misclassification. From the line plot in blue in Fig. 4(c), it can be seen that while the equal error rates are around 5% at high SNRs, it deteriorates to 33% when the SNR is -10 dB.

Our proposed framework uses a held-out set of unseen noisy speech data and non-speech data to retune the Gabor filters. The aim of such a retuning process using mismatched noisy speech and non-speech is to effectively transform the sensory mapping, so as to emphasize the specific regions of the modulation space that are *uniquely* indicative of the presence or absence of speech as reflected by the clean speech and non-speech GMM models, Φ_s and Φ_n respectively. The held-out set of noisy speech in this work was created using held-out data from non-speech classes as additive noise to speech. We hypothesize that by using a diverse held-out noisy speech set (with *mismatched* noise distortions relative to the test data), we can obtain an adapted setup that can scale to other novel and adverse mismatched conditions. The efficacy of this adaptation process is ultimately reflected in how separable the mismatched noisy speech and noise are in terms of LLR values.

Given that our framework employs parameterized modulation filters such as Gabors, it offers us a powerful nonlinear mechanism for retuning the sensory mapping by taking advantage of the degrees of freedom afforded by Gabor filters. Effectively, these degrees of freedom include gain ($\alpha_{\omega,\Omega}$), bandwidth ($\sigma_{t_{\omega\Omega}}, \sigma_{f_{\omega\Omega}}$) and orientation ($\theta_{\omega,\Omega}$); which conform to the nature of attention-driven adaptation observed in spectrotemporal modulation encoders in the mammalian auditory pathway [8], [44]. Next, we outline details of how Gabor filters are retuned.

A. Adaptation Procedure

The adaptation procedure requires an objective measure of speech/non-speech discriminability. In the current work, we use a d-prime measure (Equation (6)) as proxy for the *risk* measure shown in the schematic in Fig. 1. The symbols μ_x and σ_x denote the mean and standard deviation respectively, of the LLR (equation (5)) values for noisy speech ($x = ns$) and non-speech ($x = n$) classes.

$$d' = \frac{\mu_{\text{ns}} - \mu_n}{\sqrt{\left(\frac{1}{2}(\sigma_{\text{ns}}^2 + \sigma_n^2)\right)}} \tag{6}$$

Using this measure, the sensory mapping is then adapted by retuning each of the Gabor parameters $\Lambda = \{\sigma_{t_{\omega\Omega}}, \sigma_{f_{\omega\Omega}}, \theta_{\omega\Omega}, \alpha_{\omega\Omega}\}$ in order to maximize the separability measure for the new held-out set. In order to constrain this search, each of the parameters is confined to a range of possible values. That is, the adapted parameters $\hat{\Lambda} = (\hat{\sigma}_{t_{\omega\Omega}}, \hat{\sigma}_{f_{\omega\Omega}}, \hat{\theta}_{\omega\Omega}, \hat{\alpha}_{\omega\Omega})$ are constrained such that $\hat{\sigma}_{t_{\omega\Omega}} \in \mathbf{S}_{\sigma_\mathbf{t}}$, $\hat{\sigma}_{f_{\omega\Omega}} \in \mathbf{S}_{\sigma_\mathbf{f}}$, $\hat{\theta}_{\omega\Omega} \in \mathbf{S}_\theta$ and $\hat{\alpha}_{\omega\Omega} \in \mathbf{S}_\alpha$. The symbols $\mathbf{S}_{\sigma_\mathbf{t}}, \mathbf{S}_{\sigma_\mathbf{f}}, \mathbf{S}_\theta$ and $\mathbf{S}_\alpha$ denote sets of narrow range of values each of the parameters can adopt. While this aids in restricting the parameter search space, it does incorporate the fact that task-driven adaptation results in marginal changes in the characteristics of individual encoders in the auditory system [45]. Despite small changes at the individual filter level, the neural ensemble operates collectively to effectively track a target amidst maskers.

With a filter bank of M filters each with four tunable parameters, the parameter search space grows exponentially with

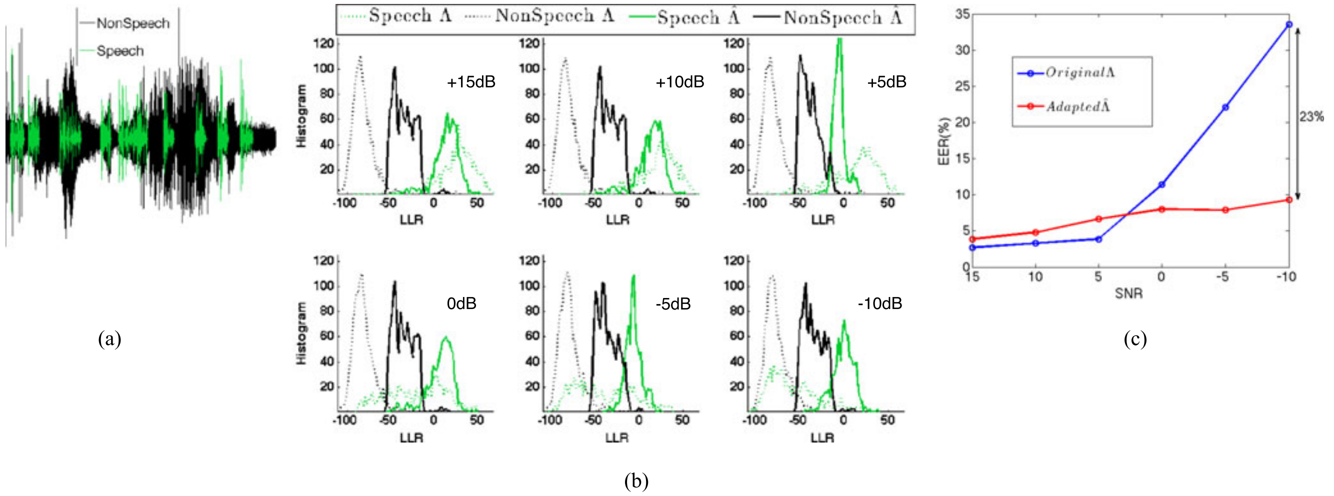


Fig. 4. a) An audio recording of 120s durations with speech in QUT StreetCity noise. b) Histogram of LLR values of speech and non-speech with speech present at different SNR, ranging from 15 to -10 dB. Dotted plots are histograms with the original filters (Λ) and solid lines represent the adapted filters $\hat{\Lambda}$. c) Equal error rate estimated using LLR values at different SNR for both default and adapted filters. The error bar on the adapted system signifies the standard deviation of the EERs obtained using different adapted filters derived from 10 independent runs of the algorithm.

cardinality of sets $\mathbf{S}_{\sigma_t}$, $\mathbf{S}_{\sigma_f}$, $\mathbf{S}_\theta$ and $\mathbf{S}_\alpha$. If each of the sets have a cardinality of C, the parameter search space is of the order $(4M)^C$. In order to sift through this large space of filter adaptations, with the goal of maximizing the d-prime measure, we employ a Genetic Algorithm (GA) as detailed next.

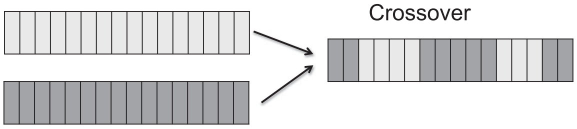
B. Genetic Algorithm

A Genetic Algorithm (GA) presents an elegant search mechanism that canvases the space of possible filter retunings in order to select an optimal solution as defined by our risk measure. The use of genetic algorithms as search heuristic for filter optimization has been more popular in image processing systems such as target detection [46], texture segmentation [47], and classification [48] problems.

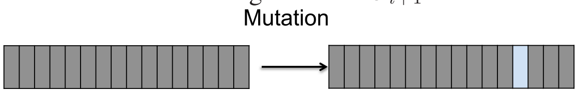
The Genetic algorithm (GA) is based on the process of evolution through natural selection and provides an effective way to search through possible members of a population P to find the fittest member. In the context of this work, each of the $(4M)^C$ bank of Gabor filters represent possible members of the population P. The d-prime estimate (equation (6)) is used as fitness measure.

Fig. 3 shows a member of population P. As can be seen each member is a $4M$ dimensioned vector, representing a *bank* of Gabor filters, with each dimension taking on values from their respective sets $\mathbf{S}_{\sigma_t}$, $\mathbf{S}_{\sigma_f}$, $\mathbf{S}_\theta$ and $\mathbf{S}_\alpha$. Using GA, we aim to find the parameter set $\hat{\Lambda}$ (fittest member) that yields maximum discriminability between speech and non-speech for a held-out data set. The algorithm works as follow:

1) Initialization: The first generation G_1 is initialized with random members of the population along with the default parameter configuration shown in Fig. 3. $G_1 = \{\Lambda, p\}$; where $p \subset P$
2) Selection: Given members of generation G_i, only the fittest members participate in generating the members of the next generation G_{i+1}. The set of retained members is denoted as $\grave{G}_i$ where $\grave{G}_i \subset G_i$.
3) Next Generation: Members of G_{i+1}, are generated in three ways.
    1) Crossover: where the members of the next generation inherit values from each of their parent members in $\grave{G}_i$ ($\approx 50\%$ from each parent) as shown in the figure below.



    1) Mutation: Members of $\grave{G}_i$ undergo changes in a limited number dimensions and take on values from appropriate $\mathbf{S}_{\sigma_t}$, $\mathbf{S}_{\sigma_f}$, $\mathbf{S}_\theta$ and $\mathbf{S}_\alpha$, to generate members of the generation G_{i+1}.



    1) Elite: Fittest members of generation $\grave{G}_i$ propagate to generation G_{i+1} without any changes.

   Thus $G_{i+1} = C(\grave{G}_i) \cup M(\grave{G}_i) \cup E(\grave{G}_i)$, where $C()$, $M()$ and $E()$ represent the crossover, mutation and elite operations respectively.
4) Stopping Criteria: Steps 2 and 3 are repeated to propagate the algorithm. The algorithm comes to a halt when the fitness of the most fit member does not change over a certain number of generations. The fittest member of the final generation is the desired $\hat{\Lambda}$.

V. Speech Activity Detection System: Experimental Methods

A. Databases

In order to test the efficacy of the proposed speech activity detection system, 3 databases were used:

1) Training set: A database of clean speech and noise was used to train the GMM statistical models of the two

TABLE I
LIST OF SPEECH AND NON-SPEECH SCENES MODELED DURING TRAINING

Speech and Non-speech Scenes	
Scene	Database
Speech	TIMIT [49]
Emergency	BBC Sound effects [50]
Office	
Impacts	
Industry	
Technology	
Transportation	
Warfare	
Water	
Weather	

classes. We used the TIMIT database [49] to train the clean speech GMM and the BBC sound effects database [50] to train the non-speech GMMs. Table I lists the scenes from the BBC sound effects database used to estimate non-speech GMM models.

1) Held-out set: A held-out set from TIMIT and BBC datasets (non-overlapping with the training set) was used for the purpose of adapting the sensory mapping process. Speech samples were corrupted additively with non-speech samples at various SNR levels ranging from 0 dB to -10 dB. A total of 800 samples each of duration 1 s, from the 2 classes were used to build this held-out set.
1) Testing sets: In order to test the system, three different corpora were used consisting of noisy speech recordings at different SNR levels with various distortions:
    1) QUT-Noise-Timit [51]: is a speech-in-noise corpus created for testing SAD systems. Noisy speech sequences are constructed using speech from TIMIT and recordings of naturally occurring noises. Car-Window-down, Street-City, Home-Kitchen and Reverb-Pool classes from the QUT database were used for testing purposes. From each of the classes 120 audio wave files, each of 60 s duration were used. The SNRs ranged from +15 dB to -10 dB with active speech proportions ranging from 25% to 75% of the audio recording. It was ensured that there was no overlap between the TIMIT data used for training the models, data used for adapting the filters and the speech data used to create noisy-speech classes in QUT.
    2) SPINE2 [52]: contains recordings in military background noise environments like aircraft carrier, humvee, office etc. All 64 recordings, on average 180 s long with 40% speech were used for testing the system. The average SNR is around 5 dB.
    3) DARPA RATS [53]: contains conversational telephone speech recorded over degraded communications channels. 250 audio files from the database, with an average duration of 700 s each were used in this work for testing the system. Audio waveforms containing regions of ”No Transmission”, which is a high amplitude static noise was excluded from scoring the SAD task as suggested by RATS guidelines. The average SNR is ≈ -2 dB. The distortion in this database is non-linear and correlated with speech.

B. Comparison With Baseline Systems

The performance of the proposed system was compared with other popular pre-trained unsupervised SAD techniques. These include single observation likelihood ratio test, with HMM based smoothing (Sohn [54]), multiple observation likelihood ratio test which uses harmonic frequency components as a primary feature (HMLRT [55]), SAD using long-term signal variability, which bases the system on the fact that speech and non-speech sounds have different variability profiles (Ghosh [56]) and the recently proposed multi-condition trained ANN system multiple source-filter model based features (Drugman [57]). It should be noted that the baseline techniques look at smaller time frames (≤ 100 ms), when compared to the proposed bio-inspired technique which seeks to capture slower temporal modulations, with an integration time of 500 ms.

C. Sensory Mapping and Model Estimation

To estimate the GMM models, spectrotemporal modulation features were extracted using the method specified in section III. Default Gabor filters $F(\omega, \Omega, t, f|\Lambda)$ were estimated at rates ω (in Hz) $\{\pm 2, \pm 4, \pm 8, \pm 16, \pm 32\}$ scales (in cycles/oct) $\{0.25, 0.5, 1, 2, 4, 8\}$, a total of 60 filters. The default parameters Λ were initialized as follows $\forall \omega, \Omega$:

$$\sigma_{t_{\omega\Omega}} = \frac{1}{2\omega}, \sigma_{f_{\omega\Omega}} = \frac{1}{2\Omega}, \theta_{\omega\Omega} = 0 \text{ and } \alpha_{\omega\Omega} = 1$$

The audio time signals were first normalized to zero mean and unit variance. The RSF representation $T(\omega, \Omega, f|\Lambda)$, was obtained by integrating the modulation space $R(\omega, \Omega, t, f|\Lambda)$ (equation (3) and (4)) over every 500 ms (frame size), with a frame shift of 10 ms. The RSF representation $T(\omega, \Omega, f|\Lambda)$ which is of dimensions $10 \times 6 \times 128$, then underwent dimensionality reduction using the TSVD based technique, to obtain a 96 dimensioned feature vector V_Λ, while ensuring that 99% of variance was retained. 48 mixture GMMs representing speech and non-speech classes with diagonal covariance were estimated using the respective 96 dimensional feature representation. Instead of a single non-speech model Φ_n as represented in equation (5), a cohort of nine non-speech GMMs were estimated, for each of the scenes in Table I, represented as $\Phi_{n_1}, \Phi_{n_2}, \ldots, \Phi_{n_9}$.

A cohort of non-speech models, were observed to work significantly better, as opposed to using a single non-speech model to represent the large variety of non-speech data. When a single non-speech model was used, it was observed that the choice of number of mixtures of the GMM model played a far too important a role. While using large number of mixtures resulted in overfitting the data, smaller number of mixtures, given the diverse nature of the data, resulted in the non-speech GMM model also encompassing the speech regions. Running the GA with a single non-speech model setup, only a marginal improvement in the d-prime measure was observed, not as significant as the one obtained with a cohort of non-speech models. The log like-

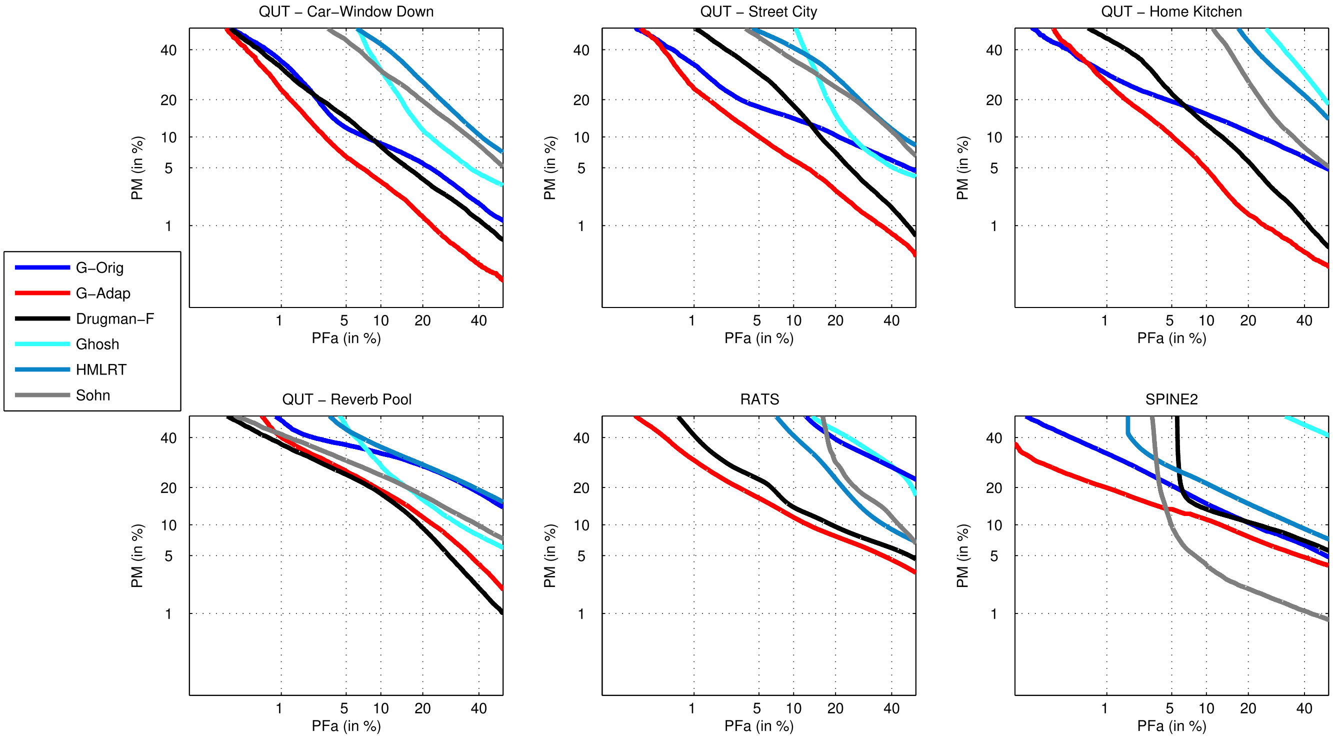


Fig. 5. DET curves using the proposed as well as the baseline algorithms for all the test databases. X axis denotes the probability of false alarm and the Y axis denotes the probability of miss.

lihood ration (LLR) required for adaptation and classification was estimated with the cohort of non-speech models as,

$$LLR = \log\left(\frac{P(V_\Lambda|\Phi_s)}{\max_{i\in\{1,2,\ldots,9\}} P(V_\Lambda|\Phi_{n_i})}\right) \qquad (7)$$

D. Sensory Mapping Adaptation

The restricted set of the values the parameters can adopt were defined as follows:

$$\mathbf{S}_{\sigma_t} = \left[\frac{1}{1.4\omega}, \frac{1}{1.6\omega}, \frac{1}{1.8\omega}, \boldsymbol{\frac{1}{2\omega}}, \frac{1}{2.2\omega}, \frac{1}{2.4\omega}, \frac{1}{2.6\omega}\right]$$

$$\mathbf{S}_{\sigma_f} = \left[\frac{1}{1.4\Omega}, \frac{1}{1.6\Omega}, \frac{1}{1.8\Omega}, \boldsymbol{\frac{1}{2\Omega}}, \frac{1}{2.2\Omega}, \frac{1}{2.4\Omega}, \frac{1}{2.6\Omega}\right]$$

$$\mathbf{S}_\theta(\text{in degrees}) = [-4.5, -3, -1.5, \mathbf{0}, 1.5, 3, 4.5]$$

$$\mathbf{S}_\alpha = [0.7, 0.8, 0.9, \mathbf{1}, 1.1, 1.2, 1.3]$$

The default parameter values are indicated in bold. With the filter-bank comprising of 60 filters and all sets having a cardinality of 7, the parameter search space is of the order 240^7. Using the held-out dataset of non-speech and low-SNR noisy speech samples, GA was run to obtain a robust set of Gabor filters represented as the parameter set $\hat{\Lambda} = (\hat{\sigma}_{t_{\omega\Omega}}, \hat{\sigma}_{f_{\omega\Omega}}, \hat{\theta}_{\omega\Omega}, \hat{\alpha}_{\omega\Omega})$.

Fig. 6 shows the performance of adaptation process in terms of d-prime of the fittest member of the population at each iteration of the Genetic algorithm. As can be seen, there is a marked increase in the d-prime measure of the fittest member in the initial few generations, with a more gradual improvement at the later generations. The d-prime measure on the held-out set was found to improve from 1.91 (Λ) to 2.34 ($\hat{\Lambda}$) over 78 generations before reaching the stopping criterion (no change in fitness over 25 generations).

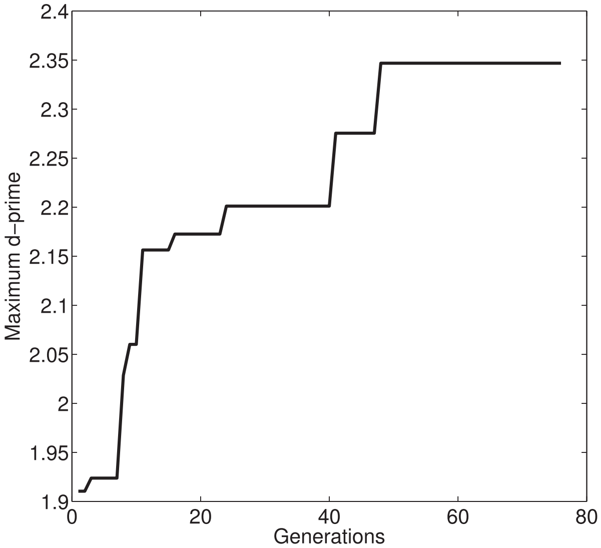


Fig. 6. Plot shows the increasing d-prime measure of the fittest member as generations propagate, stopping at generation 78.

Further, in order to establish the statistical significance of the adaptation process over the default system, the same process was repeated 10 times to obtain 10 different sets of retuned filters. In the following section, we will compare the systems derived from these independent runs with the original (unadapted) system to show statistically significant improvements across different SNRs. A t-test is performed to probe the null hypothesis that the EER values from different runs of the adaptation process are from a distribution with the mean equal to the EER of the default system (G-Orig).

E. Sensory Mapping Adaptation Versus Model Adaptation

In the proposed framework, the clean speech and non-speech models are kept fixed and are used to provide feedback to adapt the sensory mapping process. We compared the performance of the proposed system with that of a system where the model is adapted. We estimated adapted speech models, using Maximum a posteriori (MAP) technique, for each of the 6 noise cases (4 classes from QUT, Darpa Rats, SPINE2) using 300 s of noisy speech data from the respective noise class. We seek to compare the performance of such a model adaptation process vis-a-vis adapting the sensory mapping process which has no access to data from the test databases.

F. Role of SNR of Held-Out Data

For our primary system, the retuning of the filters was performed using low SNR mismatched speech data (≤ 0 dB). We also investigated how the effectiveness of the adaptation process varied on using a held-out set at different SNR levels (either higher, matched or lower than the test condition). In order to do so, first, the noisy speech component of the adaptation data was restricted to one of these SNR values (in dB), $\{15, 10, 5, 0, -5, -10\}$. Then using GA, a different set of retuned Gabor filter parameters were estimated for each of these held-out adaptation datasets. The aim here is to employ each of these 6 retuned filters on test data at different SNR, in order to study the efficiency of the adaptation process in relation to SNR of the data used for adaptation and SNR of the test data.

VI. Speech Activity Detection System: Results

A. Equal Error Rate

Fig. 7 shows the performance of all methods in terms of Equal Error Rate (EER) at various SNRs for data from the QUT database. Test data includes 3 types of additive noise as well as the reverberation case. The performance on using the original Gabor filters (pre-adaptation parameters Λ) is denoted as G-Orig and the retuned filters are denoted as G-Adap (post-adaptation $\hat{\Lambda}$). As can be seen in the bar plot, the adapted sensory mapping setup G-Adap shows a marked improvement over the default process G-Orig. The improvement is particularly significant in the low SNR conditions. The adapted process also outperforms other baseline methods by significant margin in very low SNR conditions of -5 and -10 dB SNR. The error bar over the adapted system signifies the standard deviation of the error rates obtained from 10 independent runs of the algorithm. A t-test comparing the default system (G-Orig) with the different adapted systems (G-Adap) confirms that all adapted systems are statistically significant ($p \leq 0.005$) for all SNR values lower than 10 dB, as shown in Fig. 7.

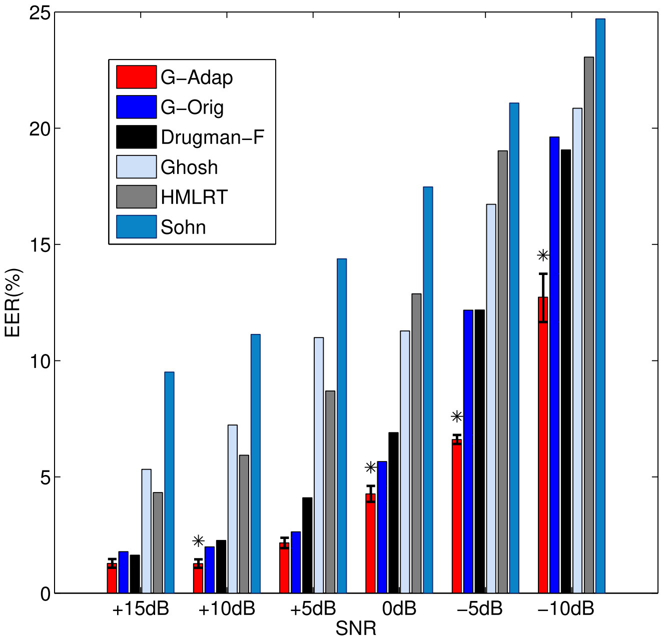


Fig. 7. Bar plot shows the average Equal Error Rate (EER) estimated using the proposed and the baseline techniques at various SNR on the 4 noise types of the QUT database. The error bar on the adapted system signifies the standard deviation of the EERs obtained on using the different adapted filters obtained during the 10 runs of the algorithm. $*$ indicates $p \leq 0.005$, on performing a t-test with the null hypothesis that the EER values from different runs of the adaptation process, is from a distribution with the mean equal to the EER of the default system (G-Orig).

Fig. 5 shows the Detection Error Trade-off (DET) curves for all three testing databases using the proposed as well as other baseline systems. For the QUT database, the DET plots are shown separately for each of the different background noise scenarios. It can be seen that on employing bio-mimetic features, G-Orig and G-Adap perform relatively well for additive noise cases, with a considerable improvement with the retuned filters. For the reverberation case, the default sensory mapping process G-Orig performs poorly, while the adapted filters G-Adap show a far superior performance. It should be noted that the adaptation was performed using additive noise and that the reverberation case represents a highly mismatched scenario. Similarly, a notable improvement can be observed with the adapted setup (G-Adap) over the default setup for the RATS database, again with non-linear noise distortions. In the case of SPINE2, the HMLRT method performs the best, though at very low probability of false alarm, the probability of miss is considerably lower on using the proposed method.

B. Spotlight on Speech and Non-Speech Unique Regions

From the EER numbers, it is evident that there is considerable enhancement of robustness on adapting the sensory mapping process, even for novel unseen test data. Except for the SPINE2 dataset, the proposed method has the lowest equal error rate when compared to other SAD systems. It is worth noting that even in the case of RATS with correlated unseen noise condi-

TABLE II
SENSORY MAPPING ADAPTATION VERSUS MODEL ADAPTATION

Database	Adaptation of Filters	Adaptation of models
QUT	8.85	**8.64**
RATS	**10.73**	23.50
SPINE2	**10.43**	23.89

tions, the adapted filters that use the original clean speech GMM still yields an equal error rate of around 10% .

To shed light on the benefits gained from the adaptation operation, we revisit the example in Fig. 4 used earlier to motivate the need for adaption in Section IV. This particular example was from the QUT Street City test database. Fig. 4(b) shows the histograms of LLR values of noisy speech and non-speech with the original (Λ) and adapted filters ($\hat{\Lambda}$). The results somewhat refute our original hypothesis that the adaptation process will repel the the speech/non-speech histograms away from each other by maximizing the discriminability between the two classes. Instead, the results reveal that the adaptation process tends to tighten the range of LLR values for both speech and non-speech especially at low SNR values, hence resulting in considerably less overlap between the speech and non-speech LLR histograms. It can be inferred from this that the adapted sensory mapping process attempts to shine the spotlight strongly on speech unique regions and certain non-speech unique regions of the RSF space. While this results in greater LLR values for both classes, it also results in less variability in terms of likelihood values when presented with noisy speech and non-speech, implying higher separability. Fig. 4(c) shows the equal error rate as a function of SNR, with the original and adapted filters. For this particular example, there is an improvement of ≈ 23 percentage points in terms of absolute EER at -10 dB SNR. The error bar signifies the standard deviation of the error rates obtained on using adapted filters estimated over 10 different runs of the algorithm on the test utterance at different SNRs.

C. Sensory Mapping Adaptation Versus Model Adaptation

The performance of the proposed system was compared to the scenario where data from a test case is available (300 s) and the models can be adapted using a Maximum a posteriori (MAP) technique (see Section V). It should be noted that for the sensory mapping adaptation system, no data from the test databases was used.

It can be seen in Table II that the proposed method (based on feature adaptation) performs significantly better over model-adapted systems when tested on the RATS and SPINE2 databases. Model adaptation leads to modest improvements over feature adaptation on the QUT database (averaged across the 4 noise types). It is important to highlight that even though 300 s of data is used to adapt the models (in case of MAP adaptation systems), the non-stationary nature of the noise (especially in the RATS and SPINE2 datasets) is insufficient to adapt the clean speech models. In such dynamic noise cases, feature-based adaptation appears to be far more effective with far less dependence on test data being available. In contrast, in relatively stationary cases such as QUT database, model-based adaptation is reasonably effective in adjusting the models; though our results indicate that feature-based adaptation performs almost similarly to MAP-adaptation.

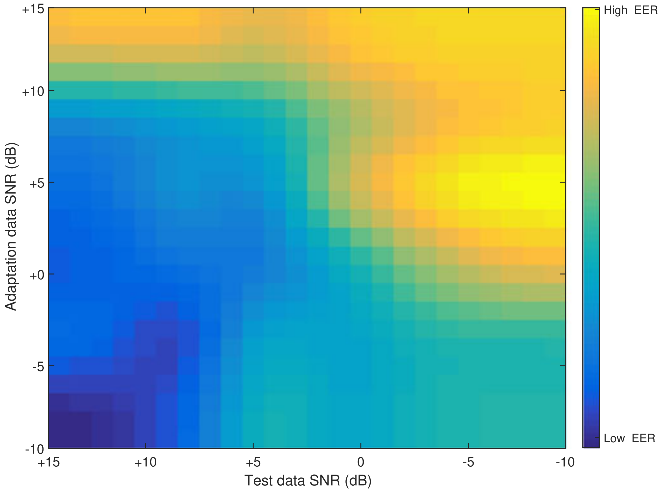


Fig. 8. A representation of the EER spread at different SNRs of QUT test data in relation to the SNR of the adaptation data used to adapt the sensory mapping process. The image was interpolated (for visualization purposes) to maintain smoothness in representation

D. Role of SNR of Adaptation Data

It should be noted that for the results discussed thus far, the retuning of the filters was performed using low SNR mismatched speech data (≤ 0 dB). Our findings indicate that such an adapted system yields significant improvement in robustness across all SNR and even in severely mismatched conditions like the reverb-pool from the QUT database and the RATS database. We next investigate the effectiveness of the adaptation process in relation to the SNR of the held-out set. Fig. 8 shows the average EER spread for the QUT database across different SNR, in relation to the SNR of the held-out adaptation data (setup explained in Section V-E).

As can be seen in the figure, even for high SNR test data, using low SNR held-out adaptation data, resulted in a more accurate system, when compared to using SNR matched adaptation data. For low SNR test data, it can be that observed that using a set of filters retuned using high SNR adaptation proves to be highly detrimental in terms of EERs. On studying the LLR values for these cases, it was observed that adaptation using just the low SNR data, resulted in far narrower histograms (Fig. 4), resulting in decreased risk of misclassification. This indicates, lower the SNR of the speech data used for adaptation, further restricted are the regions of the RSF space that are emphasized by the feedback from the models. This leads to decreased variability in LLR values; irrespective of the SNR of the test data.

E. Effectiveness of the Gabor Parameters

The parameters that were retuned were the gain $\alpha_{\omega,\Omega}$, bandwidth of the Gaussians along time $\sigma_{t_{\omega\Omega}}$ and frequency $\sigma_{f_{\omega\Omega}}$

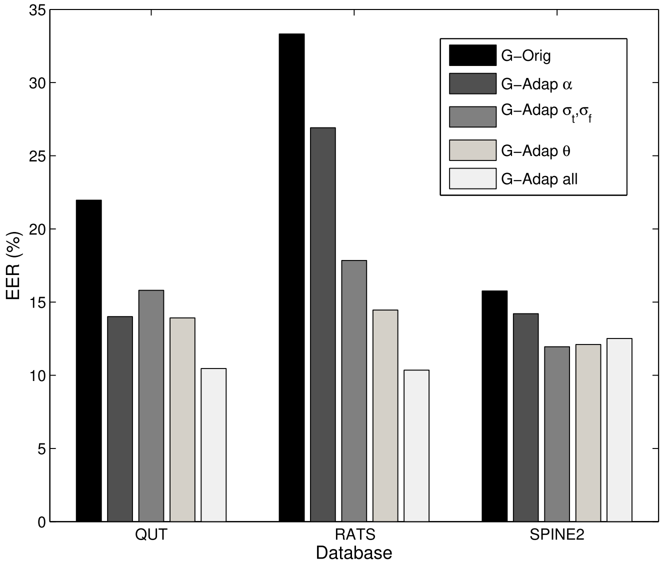


Fig. 9. Bar plot shows the Equal Error Rate (EER) estimated with the i) Original filters ii) retuning just the gain parameter α iii) parameters $\sigma_t \sigma_f$ iv) parameter θ v) retuning all parameters $\hat{\Lambda}$

and the orientation $\theta_{\omega,\Omega}$ parameters. In order to explore the effectiveness of retuning each of these three parameters, each of the parameters were retuned using GA in isolation, while keeping the other two parameters fixed at their default values. Fig. 9 shows the EER obtained for each of the three databases on employing the original default filters (*G-Orig*) and the adapted filters. It can be seen that for all 3 databases, just retuning one of the parameters also leads to an improvement in the EER. In fact for the SPINE2 database, the filters with just the bandwidths retuned perform best. For the QUT and DARPA RATS databases, retuning all three parameters (*G-Adap All)* perform best, while just retuning the orientation parameter $\theta_{\omega\Omega}$ performs next best in terms of EER.

VII. Discussion

In this work, we sought to leverage some of the key processes observed in the mammalian auditory system for the task of robust speech activity detection. We first proposed the use of two-dimensional parameterized Gabor filters to form the core of a high-dimensional adaptable feature extraction process. We then developed a mechanism using Genetic algorithm to retune the Gabor filters, driven by the feedback from clean speech and non-speech models for a held-out set of mismatched examples. The proposed discriminatory non-linear setup differs significantly from the linear transformation based setup explored in [58].

We hypothesized that such an adapted system will scale for other unseen conditions and validated it by showing marked improvement in performance of the system when tested in novel adverse conditions. It should be noted that even with a small restricted set of adaptable parameter space, there is a significant improvement in robustness on adapting the Gabor filters; thereby highlighting the importance of the joint spectrotemporal modulation space and the effectiveness of task driven adaptation within that space.

As has been emphasized earlier, the clean speech and the cohort non-speech models in this work serve as fixed statistical representations. They are then used to provide feedback for adapting the bottom-up process to enhance robustness with respect to these models. It is in this detail that the proposed methodology differs from other methods like model adaptation or multi-condition training based techniques that seek to enhance robustness of SAD system. While there have been relatively fewer efforts that incorporate such feedback-driven adaptation of the sensory mapping processes in audio processing applications, a host of such techniques have been explored in the field of machine vision [59]–[67], especially for tasks like visual object detection in complex scenes.

For instance, in [65], the popular biologically inspired HMAX model [68] of bottom-up image processing for object recognition, is first cast in probabilistic terms to form a bayesian network. Feedback as belief propagation through the bayesian network, is used to modulate the HMAX like image processing setup in line with the behavioral goals. In [67] it is shown that deep belief networks can be employed to perform hierarchical probabilistic inference and an example of top-down inference is demonstrated using face recognition data. In [69], top-down saliency is integrated into a deep network based setup for object recognition. Similar to this paper, it was shown that summary statistics of the objects in isolation (analog to clean speech model) can be used to modulate the saliency response of the network depending on the discriminatory power of the underlying features. This discriminatory power of the network is shown to be enhanced over multiple layers of the network. In [60], [61], [64], which deal with detecting visual objects in robotic applications, *long-term memory* representations of multiple objects are first learnt during the training phase. When presented with an object query, relevant top-down biases with respect to the object model are applied to enhance the ability of the system to identify the queried object.

In conclusion, in this work we have outlined a framework that leverages task driven adaptation successfully for the purpose of highly robust speech activity detection. Using a parameterized Gabor filter based setup, we have illustrated the effectiveness of adaptation in the spectrotemporal modulation space, driven by feedback from the higher level models. While in this work, the adaptation process is incorporated prior to testing, using a held-out data set, as future work, we plan to explore rapid on the fly unsupervised adaptation in novel unseen conditions, biased by the feedback from the statistical representations. In the mammalian system, it has also been observed that in a task driven setting, along with sensory mapping adaptation, there is also modulation of the cognitive and decision making areas of the brain [70]–[72]. We plan to extend the proposed framework to also leverage this aspect of the mammalian auditory system. In the context of this work, it would imply adapting the models themselves in tandem with the sensory mapping adaptation, in a complementary manner.

REFERENCES

[1] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005.

[2] J. J. Eggermont, "Between sound and perception: Reviewing the search for a neural code," vol. 157, nos. 1–2, pp. 1–42, 2001.

[3] A. N. Popper and R. R. Fay, Eds., *The Mammalian Auditory Pathway: Neurophysiology* (ser. Springer handbook of auditory research), vol. 2. New York, NY, USA: Springer-Verlag, 1992.

[4] 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. 2010.

[5] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, 2005.

[6] I. Nelken, "Processing of complex stimuli and natural scenes in the auditory cortex," *Current Opinion Neurobiol.*, vol. 14, no. 4, pp. 474–480, 2004.

[7] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Adaptive changes in cortical receptive fields induced by attention to complex sounds," *J. Neurophysiol.*, vol. 98, no. 4, pp. 2337–2346, 2007.

[8] P. Yin, J. B. Fritz, and S. A. Shamma, "Rapid spectrotemporal plasticity in primary auditory cortex during behavior," *J. Neurosci.*, vol. 34, no. 12, pp. 4396–4408, Mar. 19, 2014.

[9] S. Shamma and J. Fritz, "Adaptive auditory computations," *Current Opinion Neurobiol.*, vol. 25, pp. 164–168, Apr. 2014.

[10] G. S. V. S. Sivaram, N. S. Krishna, N. Mesgarani, and H. Hermansky, "Data-driven and feedback based spectro-temporal features for speech recognition," *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 957–960, Nov. 2010.

[11] S. K. Nemala, K. Patil, and M. Elhilali, "Recognizing the message and the messenger: Biomimetic spectral analysis for robust speech and speaker recognition," *Int. J. Speech Technol.*, vol. 16, pp. 313–322, 2012.

[12] D. J. Klein, P. Konig, and K. P. Kording, "Sparse spectrotemporal coding of sounds," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 7, pp. 659–667, 2003.

[13] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Commun.*, vol. 53, no. 5, pp. 736–752, 2011.

[14] J.-T. Huang, J. Li, and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4989–4993.

[15] S.-Y. Chang and N. Morgan, "Robust CNN-based speech recognition with Gabor filter kernels," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 905–909.

[16] T. Gramss, "Word recognition with the feature finding neural network (FFNN)," in *Proc. IEEE Workshop Neural Netw. Signal Process.*, 1991, pp. 289–298.

[17] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *INTERSPEECH*, pp. 2573–2576, 2003.

[18] T. N. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8614–8618.

[19] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 920–930, May 2006.

[20] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 2519–2523.

[21] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the DARPA RATS program." in *INTERSPEECH*, pp. 3497–3501, 2013

[22] M. Elhilali and S. A. Shamma, "Adaptive cortical model for auditory streaming and monaural speaker separation," in *Workshop on Speech Separation and Comprehension in Complex Acoustic Environments by Humans and Machines*, 2004.

[23] A. J. Simpson, "Probabilistic binary-mask cocktail-party source separation in a convolutional deep neural network," *arXiv preprint arXiv:1503.06962*, 2015.

[24] M. Slaney, T. Agus, S.-C. Liu, M. Kaya, and M. Elhilali, "A model of attention-driven scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 145–148.

[25] K. Patil and M. Elhilali, "Attentional mechanisms for recognizing acoustic scene," in *ARO Mid-Winter Meeting, Assoc. Res. Otolaryngologists*, vol. 36, New Jersey, 2013.

[26] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. 22nd Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 1096–1104.

[27] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: The biological bases of musical timbre perception," *PLoS Comput. Biol.*, vol. 8, no. 11, Nov. 2012, Art. no. e1002759.

[28] K. Patil and M. Elhilali, "Goal-oriented auditory scene recognition," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 2510–2513.

[29] M. Carlin and M. Elhilali, "A computational model of optimal adaptive changes in auditory cortical receptive fields during behavioral tasks," in *Comput. Syst. Neurosc. Meeting, COSYNE14*, Utah, 2014.

[30] P. Assmann and Q. Summerfield, "The perception of speech under adverse acoustic conditions," in *Speech Processing in the Auditory System*, vol. 18. Berlin, Germany: Springer, 2004, pp. 231–308.

[31] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2000.

[32] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention—Focusing the searchlight on sound," *Current Opinion Neurobiol.*, vol. 17, no. 4, pp. 437–455, 2007.

[33] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[34] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[35] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 483–487.

[36] X.-L. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *Proc. IEEE Int. Conf.Acoust., Speech Signal Process.*, 2013, pp. 853–857.

[37] X.-L. Zhang, "Unsupervised domain adaptation for deep neural network based voice activity detection," in *Proc. Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 6864–6868.

[38] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1990.

[39] D. L. Wang and G. J. Brown, Eds.,*Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ, USA: IEEE Press/Wiley, 2006.

[40] H. L. Hawkins, T. A. McMullen, and R. R. Fay, *Auditory Computation*. New York, NY, USA: Springer, 2012, vol. 6.

[41] T. Ezzat, J. V. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 506–509.

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

[43] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM. J. Matrix Anal. Appl.*, vol. 21, pp. 1253–1278, 2000.

[44] J. Fritz, M. Elhilali, and S. Shamma, "Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex," *Hearing Res.*, vol. 206, nos. 1/2, pp. 159–176, 2005.

[45] M. Elhilali, J. B. Fritz, T. S. Chi, and S. A. Shamma, "Auditory cortical receptive fields: Stable entities with plastic abilities," *J. Neurosci.*, vol. 27, no. 39, pp. 10372–10382, 2007.

[46] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection using evolutionary Gabor filter optimization," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 125–137, Jun. 2005.

[47] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters." *Pattern recognition*, vol. 24, no. 12, pp. 1167–1186, 1991.

[48] M. Afshang, M. S. Helfroush, and A. Zahernia, "Gabor filter parameters optimization for texture classification based on genetic algorithm," in *Proc. 2nd Int. Conf. Mach. Vision*, 2009, pp. 199–203.

[49] J. S. Garofolo *et al.*, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA. Linguistic Data Consortium, 1993.

[50] *The BBC Sound Effects Library Original Series*. [Online]. Available: http://www.soundideas.com, 2006.

[51] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 3110–3113.

[52] A. Schmidt-Nielsen *et al.Speech in Noisy Environments (SPINE2) Part 3 Audio LDC2001S08*. Philadelphia, PA, USA: Linguistic Data Consortium, 2001.

[53] K. Walker, Kevin and S. Strassel, “The rats radio traffic collection system.” in *Odyssey*, pp. 291–297, 2012.

[54] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[55] L. N. Tan, B. J. Borgstrom, and A. Alwan, “Voice activity detection using harmonic frequency components in likelihood ratio test,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2010, pp. 4466–4469.

[56] P. K. Ghosh, A. Tsiartas, and S. Narayanan, “Robust voice activity detection using long-term signal variability,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 600–613, Mar. 2011.

[57] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, “Voice activity detection: Merging source and filter based information,” *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 252–256, Feb. 2016.

[58] K. Patil and M. Elhilali, “Task-driven attentional mechanisms for auditory scene recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 828–832.

[59] D. B. Walther and C. Koch, “Attention in hierarchical models of object recognition,” *Prog. Brain Res.*, vol. 165, pp. 57–78, 2007.

[60] Y. Yu, G. K. Mann, and R. G. Gosine, “An object-based visual attention model for robotic applications,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 5, pp. 1398–1412, Oct. 2010.

[61] Y. Yu and R. G. Gosine, “Target tracking for moving robots using object-based visual attention,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 2902–2907.

[62] A. L. Rothenstein and J. K. Tsotsos, “Attention links sensing to recognition,” *Image Vision Comput.*, vol. 26, no. 1, pp. 114–126, 2008.

[63] F. H. Hamker and J. Worcester, “Object detection in natural scenes by feedback,” in *Biologically Motivated Computer Vision*. New York, NY, USA: Springer, 2002, pp. 398–407.

[64] Y. Sun, “Hierarchical object-based visual attention for machine vision,” Ph.D. dissertation, School Informat., Univ. Edinburgh, Edinburgh, U.K., 2003.

[65] S. Dura-Bernal, T. Wennekers, and S. L. Denham, “Top-down feedback in an HMAX-like cortical model of object perception based on hierarchical Bayesian networks and belief propagation,” *PloS one*, vol. 7, no. 11, 2012, Art. no. e48216.

[66] F. H. Hamker, “Modeling feature-based attention as an active top-down inference process,” *BioSystems*, vol. 86, no. 1, pp. 91–99, 2006.

[67] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.

[68] M. Riesenhuber and T. Poggio, “Hierarchical models of object recognition in cortex,” *Nature Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[69] S. Han and N. Vasconcelos, “Object recognition with hierarchical discriminant saliency networks,” *Frontiers Comput. Neurosci.*, vol. 8, 2014, Art. no. 109.

[70] J. B. Fritz, S. V. David, S. Radtke-Schuller, P. Yin, and S. A. Shamma, “Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex,” *Nature Neurosci.*, vol. 13, no. 8, pp. 1011–1019, Aug. 2010.

[71] B. G. Shinn-Cunningham, “Object-based auditory and visual attention,” *Trends Cogn. Sci.*, vol. 12, no. 5, pp. 182–186, 2008.

[72] C. Alain and S. R. Arnott, “Selectively attending to auditory objects,” *Frontiers Biosci.: A J. Virtual Library*, vol. 5, pp. D202–D212, Jan. 1, 2000.

Ashwin Bellur received the M.S. degree from the Indian Institute of Technology Madras, Chennai, India, in 2013. He is currently working toward the Ph.D. degree in electrical and computer engineering with the Laboratory for Computational Audio Perception, Johns Hopkins University, Baltimore, MD, USA. From 2009 to 2013, he was a Research Assistant with the Indian Institute of Technology Madras, where he worked on text-to-speech synthesis and music-information-retrieval-based research problems. His research interests include computational neuroscience, auditory scene analysis, speech processing, and machine learning.

Mounya Elhilali (S’00–M’08–SM’16) received the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 2004. She is an Associate Professor with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD. She directs the Laboratory for Computational Audio Perception and is affiliated with the Center for Speech and Language Processing. Her research examines the computational and neural bases of sound and speech perception in complex acoustic environment and looks at problems of auditory scene analysis, cognitive control, and sound perception in multisource settings. Applications of her work span speech systems, content analysis technologies, and medical applications. She received the National Science Foundation CAREER Award and the Office of Naval Research Young Investigator Award. She is the Charles Renn Faculty Scholar.