SOUND OBJECT RECOGNITION

by

Kailash Patil

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

December, 2013

© Kailash Patil 2013

All rights reserved

Abstract

Humans are constantly exposed to a variety of acoustic stimuli ranging from music and speech to more complex acoustic scenes like a noisy marketplace. The human auditory perception mechanism is able to analyze these different kinds of sounds and extract meaningful information suggesting that the same processing mechanism is capable of representing different sound classes. In this thesis, we test this hypothesis by proposing a high dimensional sound object representation framework, that captures the various modulations of sound by performing a multi-resolution mapping. We then show that this model is able to capture a wide variety of sound classes (speech, music, soundscapes) by applying it to the tasks of speech recognition, speaker verification, musical instrument recognition and acoustic soundscape recognition.

We propose a multi-resolution analysis approach that captures the detailed variations in the spectral characterists as a basis for recognizing sound objects. We then show how such a system can be fine tuned to capture both the message information (speech content) and the messenger information (speaker identity). This system is shown to outperform state-of-art system for noise robustness at both automatic speech recognition and speaker verification tasks.

The proposed analysis scheme with the included ability to analyze temporal modulations was used to capture musical sound objects. We showed that using a model of cortical processing, we were able to accurately replicate the human perceptual similarity judgments and also were able to get a good classification performance on a large set of musical instruments. We also show that neither just the spectral feature or the marginals of the proposed model are sufficient to capture human perception. Moreover, we were able to extend this model to continuous musical recordings by proposing a new method to extract notes from the recordings.

Complex acoustic scenes like a sports stadium have multiple sources producing sounds at the same time. We show that the proposed representation scheme can not only capture these complex acoustic scenes, but provides a flexible mechanism to adapt to target sources of interest. The human auditory perception system is known to be a complex system where there are both bottom-up analysis pathways and top-down feedback mechanisms. The top-down feedback enhances the output of the bottom-up system to better realize the target sounds. In this thesis we propose an implementation of top-down attention module which is complimentary to the high dimensional acoustic feature extraction mechanism. This attention module is a distributed system operating at multiple stages of representation, effectively acting as a retuning mechanism, that adapts the same system to different tasks. We showed that such an adaptation mechanism is able to tremendously improve the performance of the system at detecting the target source in the presence of various distracting background sources.

Primary Reader: Dr. Mounya Elhilali

Secondary Reader: Dr. Andreas G. Andreou

Acknowledgments

I am immensely thankful to my advisor and mentor Dr. Mounya Elhilali, for providing encouragement and freedom to pursue my research without any restrictions. She provided support and advise not only for this thesis but for any other research problems that I was interested in and for my career as well. For that I would be eternally grateful to her.

I am also grateful to Sridhar Nemala, for the collaborative work and for all his input and discussions without which some parts of this thesis would not have been possible. I am also thankful to Michael Carlin for all the support, discussions and fun projects that we were involved in right from my first days at Johns Hopkins. The other members who were involved in my work are Samuel Thomas and Sriram Ganapathy who have also provided guidance throughout my career. I would also like to thank the other members of my lab who made life at Johns Hopkins very enjoyable.

On a personal level, I would like to thank my parents Bhagawantraya, Sadhana and family members, Seema, Ravi and Dakshayani for all the love, encouragement and support during my stay at Johns Hopkins. I would also like to thank Anasuya, who has been a pillar in life, and enduring with me throughout the 5 years, the hardships and ups and downs of graduate life.

Finally, I would like to thank my committee members Andreas Andreou and Sanjeev Khudanpur for their comments and suggestions on my dissertation work.

Dedication

This thesis is dedicated to Bhagawantraya, Sadhana, Anasuya and Seema.

Contents

List of Tables

List of Figures

Chapter 1

Principles of Sound Recognition

1.1 Sound Objects

1.1.1 Introduction

We experience auditory stimuli persistently, and the experience of analyzing and making sense of auditory stimuli is second nature to us. However defining auditory objects is a non-trivial task. Identity of sound objects are not necessarily tied to the source of the sound. For example, in an orchestra, the entire violin section can sound as one object. Hence identity of sound objects is more closely related to the perception of sound. Sound objects can be very fine grained objects like single notes played by a musical instrument, single phoneme spoken by a human or a single burst by a firearm. Sound objects can also be very complex like a musical performance by a band, cocktail sound of multiple speakers or the cockaphony of a busy street.

Humans, however, are very adept at analysing all these various sound objects effortlessly. Moreover, humans can adapt to different conditions and pay attention to targets of importance. For example, they can pay attention to a friend's speech in a noisy street, or they can identify their dogs bark amongst other barks in a park.

The mammalian brain relies on the same sensory pathway to carry the auditory information from all kinds of complex acoustic signals (speech, music, complex scenes etc) to the higher cognition areas. Thus similar transformations occur on all incoming signals to convert the sound pressure waveform into intermediary signals and finally meaninful perceptions. These transformations are known to be based on analysis the modulations of sound in a distributed, multi-resolution and high dimensional mappings. In this thesis we propose a model for sound processing based on the physiological observations and show how this model is capable of identifying speech sounds, musical instruments and complex acoustic scenes.

1.1.2 Object Recognition Literature

Previous attempts at recognizing sound objects consider representations that are specifically designed for the kind of sound objects considered. In this section we summarize a few relevant models proposed for musical timbre, speech representation and acoustic scene classification.

The sounds of different musical instruments form uniquely identifiable sound objects. This identity is given to the sound object by the timbre of that instrument. Although humans are able to identify timbre intuitively, the underlying dimensions of timbre are hard to pinpoint. Measures of spectral shape have been proposed as basic dimensions of timbre (e.g., formant position for voiced sounds in speech, sharpness, and brightness) [2, 3]. But timbre is not only in the spectrum, as changes of amplitude over time, the so-called temporal envelope, also have strong perceptual effects [4, 5]. To identify the most salient timbre dimensions, statistical techniques such as multidimensional scaling have been used; where perceptual differences between sound samples were collected and the underlying dimensionality of the timbre space inferred [6, 7]. These studies suggest a combination of spectral and temporal dimensions to explain the perceptual distance judgments, but the precise nature of these dimensions varies across studies and sound sets [8, 9].

Recent studies have actively explored the neural underpinnings of timbre perception. Correlates of timbre dimensions suggested by multidimensional scaling studies have been observed using event-related potentials(ERPs) [10]. Other imaging studies have attempted to identify the neural substrates of natural sound recognition, by looking for brain areas that would be selective to specific sound categories, such as voice-specific regions in secondary cortical areas [11, 12] and other sound categories such as tools [13] and musical instruments [14].

To bridge this gap, we investigate how cortical processing of spectro-temporal modulations can subserve both sound source recognition of musical instruments and perceptual timbre judgments. Specifically, computational models of cortical receptive fields are shown to be suited to classify a sound source from its evoked neural activity, across a wide range of instruments, pitches and playing styles, and also to predict accurately human judgments of timbre similarities.

Given that humans are quite adept at communicating even at relatively high levels of noise [15], numerous noise robustness approaches in the literature focused on bringing biological intuition and knowledge into front-end feature extraction (e.g. Mel cepstral analysis, perceptual linear prediction, RASTA filtering) [16]. Additional techniques were also proposed to address mismatches due to stationary or slow-varying noise/channel; such as spectral subtraction (SS) [17], log-DFT mean normalization (LDMN), long-term log spectral subtraction (LTLSS), cepstral mean normalization (CMN) [18], and variance normalization [19]. Overall, state-of-the-art systems rely on a combination of auditory motivated front-end schemes augmented with feature normalization techniques [20]. Such schemes are often augmented with speech enhancement front-ends in order to tackle mismatch conditions [21, 22]. Nonetheless, current techniques remain quite limited in dealing with various classes of distortions, including nonstationary noise sources, reverberant noises and slowly-varying channel conditions.

Attempts at automatic acoustic event and scene classification have typically followed the path of extracting short term features from waveforms and learning the statistics of these features to later classify an unknown example. Mel Frequency

Cepstral Coefficients (MFCC), filterbank energies or Perceptual Linear Prediction coefficients (PLP) have been popularly used as features for this task [23–25]. They are often complimented with other low level features like zero crossing rate, short time energy, spectral flux, pitch, brightness and bandwidth [24, 26, 27] or are transformed to account for long term statistics [28]. Though short term spectral attributes coupled with low-level features have been quite successful in a number of applications, studies have also shown that they are limited in capturing the full range of information relevant for acoustic scene recognition; and that joint local modulations in energy along both time and frequency are able to better capture the qualities of acoustic scenes [29]. This rich modulation space builds on neurophysiological studies in the mammalian auditory system indicating that neurons at the level of auditory cortex respond to local joint spectral and temporal modulation in the signal [30]. This biological analysis can be viewed as mapping sound onto a high dimensional feature space which captures the detailed variations of the spectral profile and its temporal variations, as a basis for representing acoustic events.

1.1.3 Proposed approach

The proposed work aims at 1) developing a neuro-computational framework for recognition of sound objects based on neurophysiological processing of auditory stimuli, 2) investigating the role of goal-directed feedback mechanisms in modulating the efficacy of this recognition process. The former is useful for a wide range of tasks like musical instrument recognition, scene classification, speaker recognition, etc. The latter aims at giving insight into mechanisms of attention and their role in adapting sensory processing to tasks or goals and changing acoustic environments. Unlike object recognition tasks in the visual system, the auditory modality presents us with a number of challenges; the least of which is the lack of a clear definition of what a sound object is. Moreover, natural and artificial visual scenes often contain a large proportion of static or slow-moving elements, auditory scenes are essentially dynamic; and are frequently present in conjunction with other masking effects which are often dynamic as well. A successful effort in tackling this problem has tremendous applications in engineering systems in the fields of prosthetics and communication aids, hearing technologies, surveillance, defense systems as well as robotics.

1.2 Biomimetic Object Recognition

One of the most remarkable feats that humans are able to perform rapidly and reliably is to recognize and understand the complex acoustic world that surrounds them. This process is a multi-faceted problem which encompasses various aspects of auditory perception. It encompasses the ability to detect, identify and classify sound objects; to robustly represent and identify these objects in multisource environments; and to guide actions and behaviors in line with complex goals and shifting acoustic soundscapes. Such capabalities could provide much needed robustness and flexibility to a number of technologies. Hence we look at a hierarchical system of object recognition that has been motivated by the neurophysiological processes in the mammalian auditory cortex.

1.2.1 Hierarchical System of Object Recognition

The hierarchical system of object recognition we propose consists of four stages as shown in Figure 1.1. The basic acoustic representation stage mimics the behaviour of the sub cortical processes and results in converting an incoming waveform into a time-frequency representation. The high dimensional acoustic representation stage mimics the behaviour of the auditory cortex and analyzes the time-frequency representation for a number of different cues. The perceptual object representation stage then collects statistics of different sound classes and behaves like the higher levels of the auditory cortex in helping to recognize objects. Finally, the attentional feedback stage adapts the high-dimensional acoustic representation stage and the object representation stage based on the task at hand. Each of these stages are described below.

1.2.1.1 Basic Acoustic Representation

The acoustic signal reaching the ear drum undergoes transformation from pressure waves in the air to waves along the cochlear membrane. This transformation distributes the energy at different frequencies to different locations on the membrane.

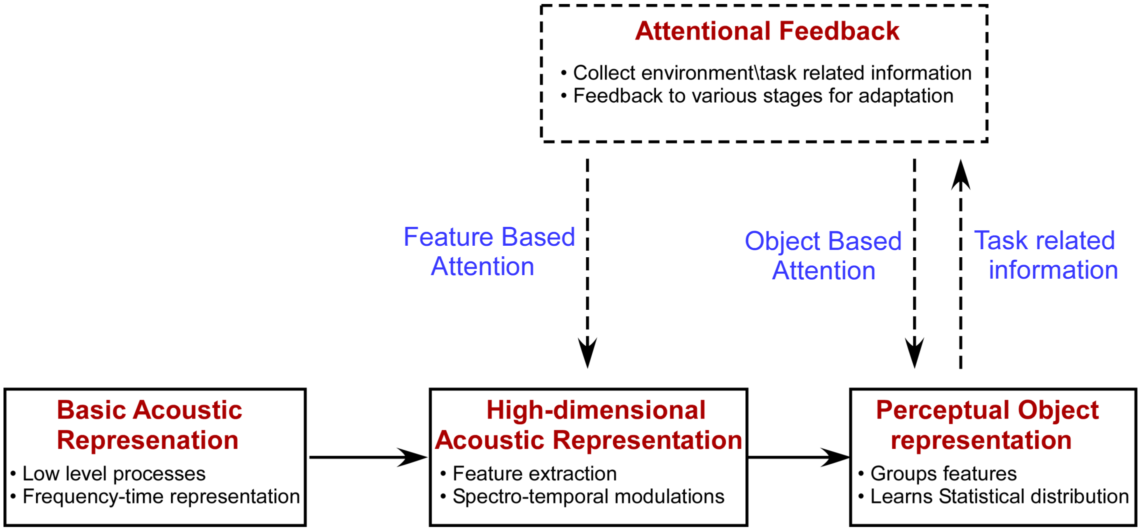
Attentional Feedback
• Collect environment\task related information
• Feedback to various stages for adaptation
Feature Based Attention
Object Based Attention
Task related information
Basic Acoustic Represenation
• Low level processes
• Frequency-time representation
High-dimensional Acoustic Representation
• Feature extraction
• Spectro-temporal modulations
Perceptual Object representation
• Groups features
• Learns Statistical distribution

Figure 1.1: Schematic of the proposed sound object recognition model. The basic and high-dimensional acoustic representation modules project the acoustic stimuli into a robust, discriminative high dimensional representation. The perceptual object recognition module collects the statistics of sound objects and learns to discriminate classes. The attentional feedback module relays back information regarding the task or environment to influence the cortical processing and object recognition modules

The frequencies are spread out on a logarithmic scale along the length of the cochlear membrane. Inner hair cells then transform these mechanical vibrations to nerve firings in the auditory nerve, which are also tonotopically organized. These stimuli then travel to the superior olivary complex, via the cochlear nucleus, where sound source localization cues, like interaural time and level differences are extracted. As these stimuli proceed to the inferior colliculus in the mid-brain, enhanced frequency selectivity, integration of localization cues and loss of phase locking to higher frequencies occurs. At this stage, the main representation is still primarily along two dimensions namely time and frequency. Details of the auditory system can be found in [31].

In the subcortical stage, the waveform is passed through a set of 128 asymmetric filters $h(t; f)$ placed uniformly on a logarithmic axis covering 5.3 octaves starting from $180Hz$. This is similar to the frequency-space transformation of the cochlear membrane. This is followed by a spectral derivative and a half wave rectification stage, which models the lateral inhibition networks in the cochlear nucleus, sharpening the frequency resolution of these filters. The mid brain processing is implemented as a short term integration with window $\mu(t; \tau) = e^{-t/\tau}u(t)$ and $\tau = 4ms$ followed by cubic root compression. These subcortical transformations can be collectively written as in Eq. 1.1 and the details of implementation can be found in [32].

$$z(t, f) = (max(\partial_f(s(t) \otimes_t h(t; f)), 0) \otimes_t \mu(t; \tau))^{\frac{1}{3}} \tag{1.1}$$

where $\otimes_t$ represents convolution with respect to time.

This resulting time-frequency representation is referred to as the auditory spectrogram.

1.2.1.2 High Dimensional Acoustic Representation

This information then reaches the primary auditory cortex (A1) via the medial geniculate body. A1 contains continuums of neurons which exhibit selectivity not just to the frequency content of sounds; but also to details of their spectral shapes and bandwidths as well as their temporal dynamics. Unlike the implicit encoding of these features in peripheral and midbrain stages, the representation at the level of A1 appears to explicitly map sounds onto a rich high-dimensional space. Here, we speculate that this multidimensional cortical representation provides a rich, potentially over-complete, basis that highlights unique aspects of sound objects; hence allowing us to identify them under their different acoustic manifestations; and even in presence of interferers.

The transformation of the stimuli in the auditory cortex is modeled as a filter-bank corresponding to a 2D affine wavelet transform, with a spectro-temporal mother wavelet, defined to be Gabor-shaped in frequency and exponential in time. Each filter is tuned ($Q = 1$) to a specific rate (ω in Hz) of temporal modulations and a specific scale of spectral modulations (Ω in cycles/octave), and a directional orientation (+ for upward and - for downward). For input spectrogram $z(t, f)$, the response of each STRF in the model is given by Eq. 1.2.

$$r_{\pm}(t, f; \omega, \Omega; \theta, \phi) = z(t, f) *_{t,f} STRF_{\pm}(t, f; \omega, \Omega; \theta, \phi) \tag{1.2}$$

where $*_{t,f}$ denotes convolution in time and frequency and θ and ϕ are the characteristic phases of the STRF's which determine the degree of asymmetry in the time and frequency axis respectively. Details of the design of the filter functions $STRF_{\pm}$ can be found in [33].

An alternative and more generalized way to analyze the spectro-temporal modulations in the spectrogram is by using 2D Gabor filters [34, 35]. Similar to the earlier model, it would give us a view of the various spectro-temporal modulations at each frequency, but would also provide us control over the gain and tuning of each filter. This would be a useful property as described later in Section 4. This set of two dimensional Gabor filters are defined as Eq. 1.3.

$$G(t, f; \sigma_t, \sigma_f) = \frac{1}{2\pi\sigma_t\sigma_f} e^{-\frac{1}{2}\left(\frac{t^2}{\sigma_t^2}+\frac{f^2}{\sigma_f^2}\right)} e^{2\pi i(\omega t+\Omega f)} \tag{1.3}$$

Where σ_t, σ_f denote the spread/tuning of the filter in time and frequency direction respectively. The response of each filter is then given by a 2d convolution with the auditory spectrogram similar to that in Eq. 1.2.

1.2.1.3 Perceptual Object Representation

The information from the primary auditory cortex then travels to higher stages in the auditory cortex where the formation and perception of the sound objects occur. Very little is known about the actual processing at these stages. In this thesis we use a diverse set of approaches to model the distribution of features from the high dimensional acoustic representation described above. These approaches have been chosen to fit the particular tasks which we were attempting to solve and will be described in detail in the following chapters.

1.2.2 Attentional Feedback

This feed forward system is augmented with feedback from higher levels to many stages as early as the periphery [31]. The functions and mechanisms of these feedbacks are relatively less well known, but they are believed to assist in task-related attention modulation [36]. Numerous experimental findings have shown that receptive fields at the level of A1 adapt their spectral and temporal properties when animals are engaged in behavioral tasks, potentially to enhance the representation of sound object of interest. These findings from ferrets [37], guinea pigs [38] and monkeys [39] have all suggested that the neural representation of sensory information can be modulated by cognitive processes such as attention in order to mediate behavior. Such findings appear at odds with psychoacoustic results indicating that auditory attention operates at a much higher-level; namely at the level of perceptual objects [40–42]. The claim is that objects are formed by integrating features from low-level cues. Top-down attention then plays a role in selecting the objects relevant to the current task. There is no unifying model of auditory attention that accounts for these findings. However, some attempts have been made to model top-down attention in the visual domain [43]. Also, there is evidence to show that similar attentional mechanisms might operate on different sensory modalities [39, 41, 44]. In Chapter 4 we propose a mechanism to allow for top-down influences at multiple stages of processing and show how this can be extremely beneficial for target detection.

1.3 Thesis Organization

The remainder of the thesis is organized as follows. In Chapter 2 we discuss a modulation based approach to noise-robust speech recognition system. In Chapter 3 we describe how the biomimetic object recognition system is able to capture the

Timbre space and an application of such a model to musical performances. In Chapter 4 we describe the proposed model of top-down attentional mechanisms and show how we are able to attend to targets in a multisource environment. Finally, in 5 we present the conclusions of this work and some possible future directions.

Chapter 2

Multi-dimensional Representations for Speech Sound Objects

2.1 Chapter Outline

Humans are quite adept at communicating in presence of noise. However most speech processing systems, like automatic speech and speaker recognition systems, suffer from a significant drop in performance when speech signals are corrupted with unseen background distortions. Here we explore the use of the proposed hierarchical system of sound object recognition (Section 1.2.1) for speech representation. However, here we focus on the information-rich spectral attributes of speech as it presents an intricate yet computationally-efficient analysis of the speech signal by careful choice of model parameters. Further, the approach takes advantage of an information-theoretic analysis of the message and speaker dominant regions in the speech signal, and defines feature representations to address two diverse tasks such as speech and speaker recognition. The proposed analysis surpasses the standard Mel-Frequency Cepstral Coefficients (MFCC), and its enhanced variants (via mean subtraction, variance normalization and time sequence filtering) and yields significant improvements over a state-of-the-art noise robust feature scheme, on both speech and speaker recognition tasks.

2.2 Introduction

Despite the enormous advances in computing technology over the last few decades, progress in the fields of automatic speech recognition (ASR) and automatic speaker verification/recognition (ASV) still faces tremendous challenges when dealing with realistic acoustic environments and signal distortions. Tackling both speech and speaker feats adds additional hurdles since information about the speaker identity and the speech message tends to be reflected in slightly distinct yet overlapping components of the speech signal. For instance, whereas formant frequencies convey crucial information about the articulatory configuration of the vocal tract, they also reveal details about speaker-specific vocal tract geometries. Yet, our brains efficiently decode the signal information pertaining to *both* speech content and speaker identity using a common front-end machinery that is quite robust even at relatively high levels of distortion and noise [45].

Mel-Frequency Cepstral Coefficients (MFCC) are a classic example of the successful influence of biological intuition onto speech technologies, making them a staple in state-of-the-art ASR and ASV systems [20, 46]. MFCCs provide a compact form of representing spectral details in the speech signal, that is motivated by both perceptual and computational considerations. They exploit the unique nature of frequency mapping in the auditory system, by warping the linear frequency axis into a nonlinear quasi-logarithmic scale. They also allow the decoupling of the speech production source and vocal tract characteristics via homomorphic filtering. In doing so, they highlight information about both the characteristics and configuration of the speech articulators that can be translated into a parametrization of both the identity of the speaker as well as the content of the speech message. While quite efficient and successful in conveying this information, features like MFCCs remain limited by their global analysis of the frequency spectrum. For instance, the first few coefficient describe details of the spectral tilt and compactness in the spectrum; but across *all* frequencies. Such broad analysis scatters information in specific frequency regions across all cepstrum coefficients.

In contrast, our knowledge of the central auditory system reveals that neurons in the auditory midbrain and primary auditory cortex exhibit a tuning to spectral details that is localized along the tonotopic axis [30,47,48]. Such neural architecture provides a detailed multi-resolution analysis of the spectral sound profile that can bear great relevance to the front-end feature schemes used in speech and speaker recognition systems. Only few studies have attempted to translate the intricate multiscale cortical processing into algorithmic implementations for speech systems, yielding some improvements for ASR tasks (in noise) albeit at the expense of great computational complexity [49, 50]. To the best of our knowledge, no similar work was done for speaker recognition.

Admittedly, translating neurophysiological strategies into compact and efficient signal processing methods comes with a number of challenges; which have often hindered the introduction of biomimetic front-ends for such complex tasks as ASR or ASV [51]. They often amount to complex and computationally-intensive mappings that are impractical to use in real systems. In the present work, we set out to devise a simple, effective, and computationally-efficient multi-resolution representation of speech signals that builds on the principles of spectral analysis taking place in the central auditory system. By carefully optimizing the choice of model parameters, the analysis constrains the signal encoding to a perceptually-relevant subspace that maximizes recognition in presence of noise while maintaining computational efficiency. Further, unlike any of the previous approaches, speech (linguistic message) and speaker (identity) dominant regions in the signal encoding are analyzed, and different parameters are defined for speech and speaker recognition tasks. By employing the same front-end processing machinery, we maintain a generic framework for speech processing that can change parameters to shift focus either towards speech content information for ASR tasks or speaker information for ASV tasks. The following section describes details of the proposed multi-resolution spectral model and motivates the choice of its parameters. Next, we describe the experimental setup and results. We finish with a discussion of the proposed analysis, and comment on potential extensions towards achieving further noise robustness.

2.3 Methods

The parameterization of speech sounds is achieved through a multistage auditory analysis that captures processing taking place at various stages along the auditory pathway from the periphery all the way to the primary auditory cortex (A1). We first describe the peripheral analysis used to obtain the 'auditory spectrogram' representation, followed by a detailed description of the cortical analysis for multi-resolution parameterization of the speech input.

2.3.1 Peripheral Analysis

The peripheral analysis of the incoming audio signal is modeled as described in Section 1.2 and is given by Equations 1.1. The outcome of this analysis is a transformation of the one-dimensional signal $s(t)$ into a time-frequency spectrogram $y(t, f)$ consisting of 128 frequency channels covering 5.3 octaves(Figure 2.1(a)). The resultant spectrogram exhibits a number of characteristics; most importantly, in preserving detailed speech information such as formant structure as well as exhibiting noise robustness qualities over conventional representations [52–54].

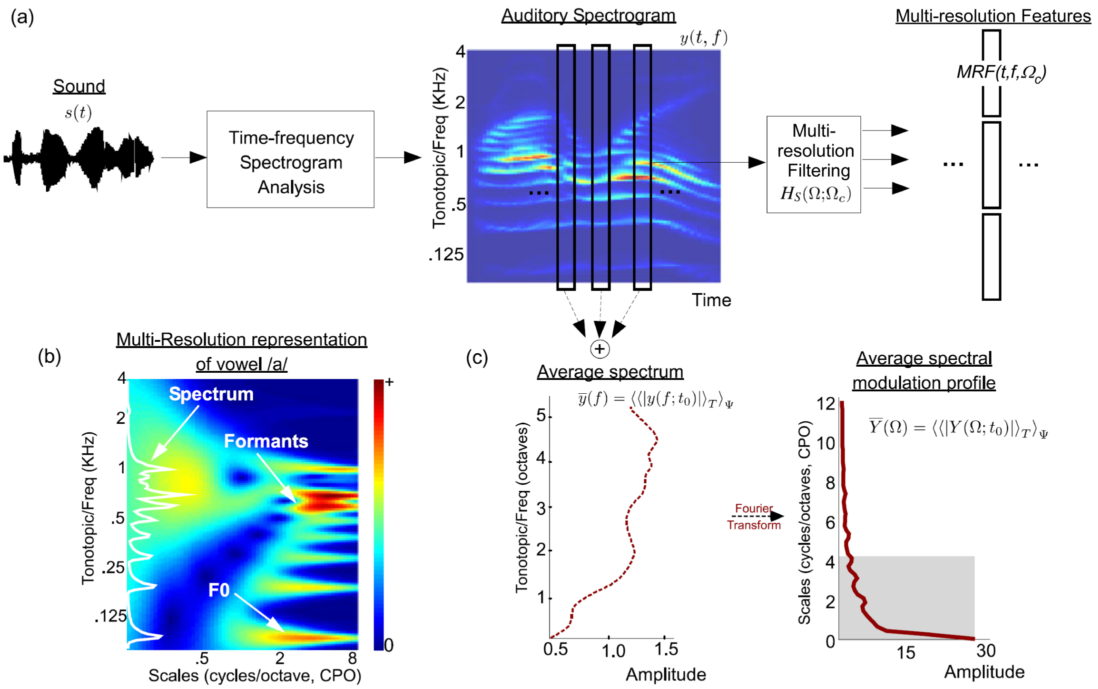
(a)
Sound
$s(t)$
Time-frequency Spectrogram Analysis
Auditory Spectrogram
$y(t, f)$
Tonotopic/Freq (KHz)
4
2
1
.5
.125
Time
Multi-resolution Filtering
$H_S(\Omega;\Omega_c)$
Multi-resolution Features
$MRF(t,f,\Omega_c)$
(b)
Multi-Resolution representation of vowel /a/
Spectrum
Formants
F0
Tonotopic/Freq (KHz)
4
2
1
.5
.25
.125
.5
2
8
Scales (cycles/octave, CPO)
+
0
(c)
Average spectrum
$\overline{y}(f) = \langle\langle|y(f;t_0)|\rangle_T\rangle_\Psi$
Tonotopic/Freq (octaves)
5
4
3
2
1
0.5
1.0
1.5
Amplitude
Fourier Transform
Average spectral modulation profile
$\overline{Y}(\Omega) = \langle\langle|Y(\Omega;t_0)|\rangle_T\rangle_\Psi$
Scales (cycles/octaves, CPO)
12
10
8
6
4
2
0
15
30
Amplitude

Figure 2.1: (a) Processing stages starting from an acoustic waveform $s(t)$ to obtain proposed features, parameterized by time t, tonotopic frequency f and spectral modulation filter parameter Ω_c. (b) Example of spectral details revealed by multi-resolution analysis for vowel /a/ (c) (left) Average auditory spectrum computed over the TIMIT corpus, $\overline{y}(f) = \langle\langle|y(f;t_0)|\rangle_T\rangle_\Psi$; (right) Average spectral modulation profile, $\overline{Y}(\Omega) = \langle\langle|Y(\Omega;t_0)|\rangle_T\rangle_\Psi$

2.3.2 Cortical Analysis

The spectrogram reveals layered information about the speech signal that is distributed over different frequency bands and varying over multiple time-constants. The next stage of processing extracts detailed information about the spectral shape in $y(t, f)$ via a bank of modulation filters operating in the Fourier domain resulting in the spectral cortical representation. The analysis mimics the spectral tuning of neurons in the central auditory pathway in which individual neurons are not only tuned to specific tonotopic frequencies (like cochlear filters); they are also selective to various spectral shapes, in particular to peaks of various widths on the frequency axis, hence expanding the cochlear one dimensional tonotopic axis onto a two-dimensional sheet [47, 55]. This analysis provides a more localized mapping of the spectral profile; that not only highlights details of bandwidth and spectral patterns in the signal but centers around the different frequency channels (Figure 2.1(b)). Mathematically, the multi-resolution spectral analysis is modeled by taking the Fourier transform of each spectral slice $y(t_0, f)$ in the auditory spectrogram and multiplying it by a modulation filter $H_S(\Omega; \Omega_c)$. The inverse Fourier transform then yields the modulation filtered version of the auditory spectrogram[1]. The spectral modulation filter $H_S(\Omega; \Omega_c)$ is defined as

$$H_S(\Omega; \Omega_c) = ({}^{\Omega}\!/_{\Omega_c})^2 e^{\left[1-({}^{\Omega}\!/_{\Omega_c})^2\right]} \quad , \quad 0 \leq \Omega \leq \Omega_{max}, \tag{2.1}$$

where Ω represents spectral modulations (or *scales*) and has units of cycles/octave (CPO), parameterizing the spectral resolution at which the auditory spectrogram is analyzed. Ω_{max} is the highest spectral modulation frequency set at 12 CPO (given the spectral resolution of 24 channels per octave).

[1] The modulation filtering is performed in real domain.

2.3.3 Choice of scales

There are two important aspects in defining the proposed multi-resolution features for a specific task (ASR or ASV): **(i)** the span of the modulation filters; and **(ii)** the distribution of filters over the chosen span. In the current study, we constrain the range of scales to less than 4 CPO, since they cover more than 90% of the entire spectral modulation energy in speech (Figure 2.1(c)) and are shown to be most crucial for speech comprehension [56]. To determine the filter distribution over the range $0-4$ CPO, we employ a judicious sampling scheme in which the modulation regions with concentrated energy are sampled more densely; while the regions with less energy are sampled more coarsely. The set of scales Ω_c is chosen by dividing the average spectral modulation profile of speech (computed over the entire train data of TIMIT corpus [57]) into equal energy regions. The average spectral modulation profile $\overline{Y}(\Omega) = \langle\langle|Y(\Omega;t_0)|\rangle_T\rangle_\Psi$ is defined as the ensemble mean of the magnitude Fourier transform of the spectral slice $y(t_0,f)$ averaged over t_0 and over all speech data Ψ. The resulting ensemble profile, shown in Figure 2.1(b), is then divided into M equal energy regions Γ_i:

$$\Gamma_i = \int_{\Omega_i}^{\Omega_{i+1}} \overline{Y}(\Omega)d\Omega, \quad \Gamma_i = \Gamma_{i+1}, \quad i = 1, ..., M-1, \tag{2.2}$$

where Ω_i and Ω_{i+1} denote the lower and upper cutoffs for k^{th} band, $\Omega_1 = 0$, and $\Omega_M = 4$.

The scheme has the dual advantage of **(i)** implicitly encoding the high energy signal components which are inherently noise robust **(ii)** sampling the given modulation space with a smaller set of scales which is important both in terms of computation complexity as well the dimensionality of the resulting feature space. Setting $M = 5$, the sampling scheme results approximately in a log-scale in the spectral modulation space, at 0.25, 0.5, 1.0, 2.0, and 4.0 CPO[2]. The output of the five spectral modulation filters for an example speech utterance is shown in Figure 2.2.

2.3.4 Encoding of speech and/vs speaker information

The speech signal, discounting the environmental and channel effects, carries information about both the underlying linguistic message and the speaker identity (Figure 2.1(b)). This information is manifested in slightly distinct yet overlapping components, and to separate these components is in general a non-trivial task. The spectral modulation filtering described above captures the overall spectral profile including formant peaks by employing broad scale filters (0.25 and 0.5 CPO) as well as narrower spectral details such as harmonic and subharmonic structures using higher resolution filters (1, 2 and 4 CPO). In order to select a set of scales (Ω_c) that are relevant for diverse tasks such as speech and speaker recognition, we analyze the mutual information (MI) between the feature variables (X) encoding various scales and the

[2] The original sampling results in spectral modulations at $\{0.18, 0.59, 1.34, 2.36, 4.0\}$ CPO.

corresponding (i) underlying linguistic message (Y_l) (ii) speaker identity (Y_s). The MI, a measure of the statistical dependence between random variables [58], is defined for two discrete random variables X and Y as:

$$I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2.3}$$

To estimate the MI, the continuous feature variables are quantized by dividing the range of observed features into cells of equal volume. To characterize the underlying linguistic message, phoneme labels from the TIMIT corpus are divided into four broad phoneme classes - the variable Y_l thus taking 4 discrete values representing the phoneme categories: vowels, stops, fricatives, and nasals. The average MI, taken as the average of the MI computed across all the frequency bands for any given scale, between the feature representations at different scales and the speech message is shown in Figure 2.3(a). In the case of speaker identity, the 'sa1' speech utterance (*She had your dark suit in greasy wash water all year*) taken from the TIMIT corpus is compared across 100 different speakers - the variable Y_s taking 100 discrete values representing the speaker identity. The average MI between different scales and speaker information is shown in Figure 2.3(b)[3].

Notice that the lower scales clearly provide significantly more information about the underlying linguistic message, while the speaker information is centered around 1

[3] The difference in MI levels between the speech message and speaker identity may be attributed to the observation that the speech signal encodes more information about the underlying linguistic message as compared to speaker information.

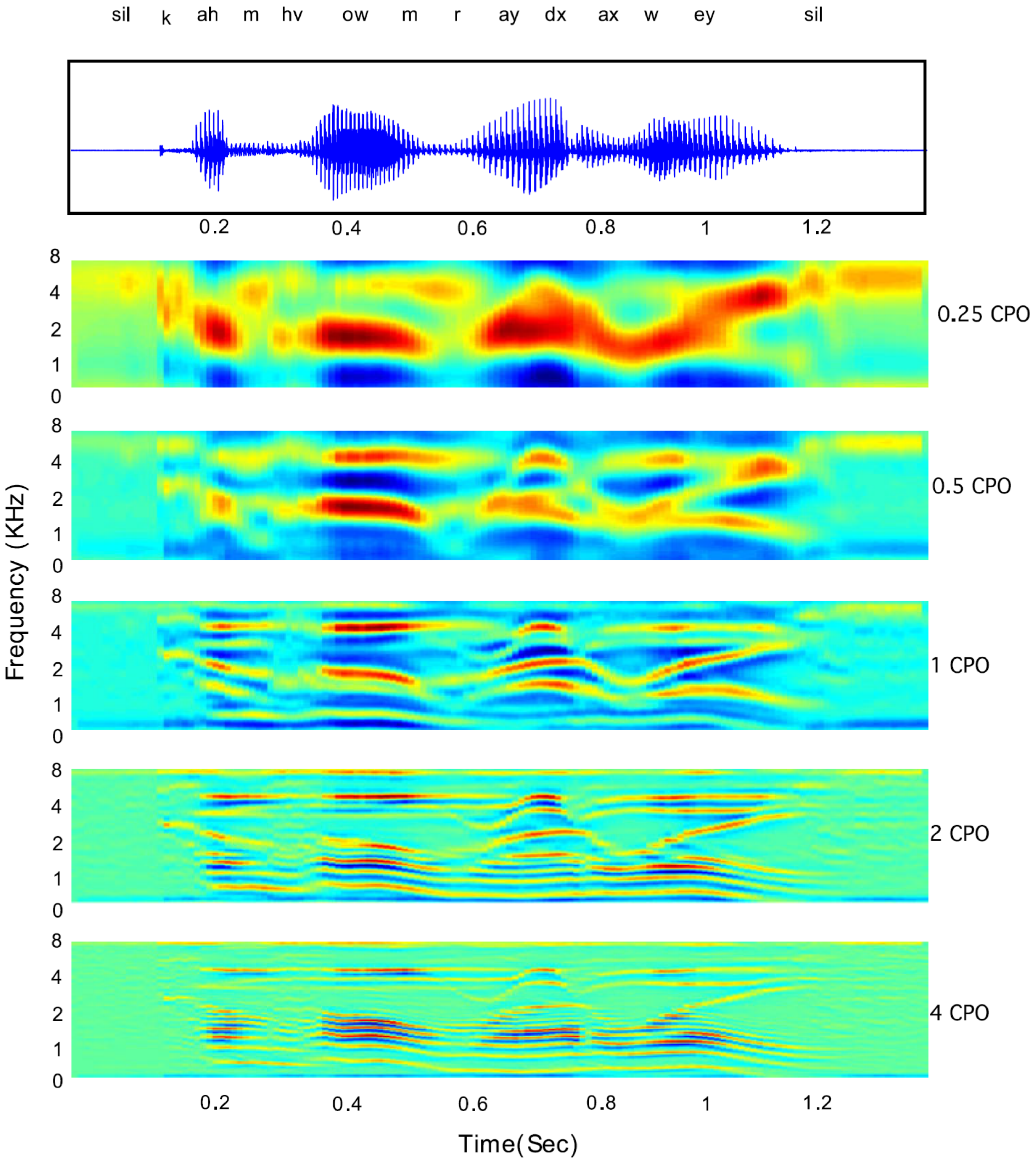

Figure 2.2: Illustration of the spectral modulation filtering at scales 0.25, 0.5, 1.0, 2.0, and 4.0 CPO for the utterance "*come home right away*" taken from TIMIT speech database. The top panel shows the time domain waveform along with the underlying phoneme label sequence.

CPO - probably highlighting the significance of overall spectral profile including formant peaks in encoding speech message and the significance of pitch or harmonically-related frequency channels in representing speaker-specific information. In order to put more emphasis on message-dominant information present in the speech signal, it is important to encode information captured by lower scales for the speech recognition task. Consequently, for the speaker recognition task it is useful to encode information captured by higher scales. Therefore, in the feature encoding for the speech recognition task we choose $\Omega_c = \{0.25, 0.5, 1.0, 2.0\}$ CPO and for the speaker recognition task $\Omega_c = \{0.5, 1.0, 2.0, 4.0\}$ CPO.

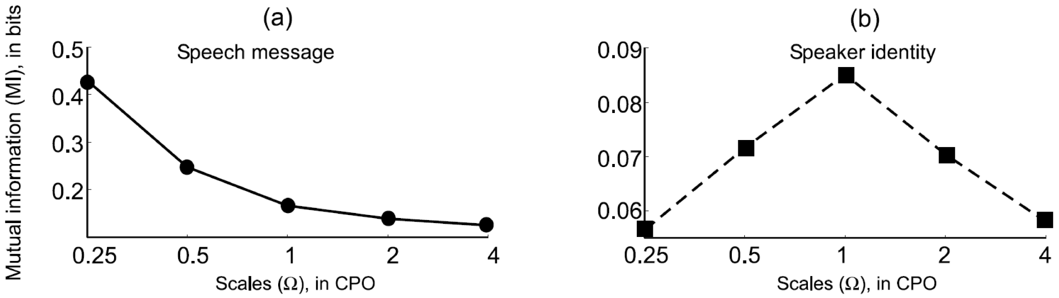

Figure 2.3: Mutual Information (MI) between feature representations encoding different scales and speech message (left panel), MI between feature representations encoding different scales and speaker information (right panel).

Finally, the filtered spectrograms (one for each scale in Ω_c) are downsampled in frequency by a factor of 4. This is achieved by integrating the 128 frequency channels into 32-bands, equally-spaced on a log-frequency axis[4]. The final multi-resolution features are defined as 128 dimensional feature vector (32 auditory frequency chan-

[4] This reduction of the spectral axis resolution did not affect the ASR/ASV performance

nels multiplied by 4 scales) at each time frame of 10 ms. An estimate of processor usage shows that computing the multi-scale modulation filtering operation on top of the auditory-inspired spectrogram increases CPU time by about 75% relative to an efficient implementation of Mel-Frequency Cesptral Coefficients.

2.4 Experimental Setup

2.4.1 Phoneme recognition setup

Speaker independent phoneme recognition experiments are conducted on TIMIT database (excluding 'sa' dialect sentences), using the hybrid Hidden Markov Model/Multilayer perceptron (HMM/MLP) framework [59–61]. The training, cross-validation and test sets consist of 3400, 296 and 1344 utterances from 375, 87 and 168 speakers respectively. 61 hand-labeled symbols of the TIMIT training transcription are mapped to a standard set of 39 phonemes along with an additional garbage class [62][5].

MLP with a single hidden layer is trained to estimate the posterior probabilities of phonemes (conditioned on the input acoustic feature vector) by minimizing the cross entropy between the input feature vectors and the corresponding phoneme target classes [63]. Temporal context is captured by training a second MLP (in a hierarchical fashion) which operates on a longer temporal context of 23 frames of

[5] It is possible to achieve higher recognition performance (in *clean* or *matched* condition) by using a larger set of 49 labels during the training and mapping to the standard set of 39 phonemes only during the scoring

posterior probabilities estimated by the first MLP [64]. Both MLPs have a single hidden layer with sigmoid nonlinearity (1500 hidden nodes) and an output layer with softmax nonlinearity (40 output nodes). The final posterior probability estimates are converted to scaled likelihoods by dividing them with the corresponding prior probabilities (unigram language model) of phonemes. An HMM with 3 states, each with equal self and transition probabilities, is used for modeling each phoneme. The emission likelihood of its each state is set to be the scaled likelihood. Finally, the Viterbi algorithm is applied for decoding the phoneme sequence. Note that the hybrid HMM/MLP system achieves better phoneme recognition performance than the standard HMM/GMM systems [65].

2.4.2 Speaker recognition setup

Text independent speaker verification experiments using Gaussian Mixture Models (GMM) are conducted on a subset of the NIST 2008 speaker recognition evaluation (SRE) [66]. In our UBM-GMM based speaker recognition system [46], the Universal Background Model (UBM) is trained with data obtained from a set of 325 speakers. In the UBM training, a total of 256 mixtures and 10 expectation-maximization iterations for mixture split are used. A total of 85 target speaker models are obtained by *maximum a posteriori* (MAP) adaptation of the UBM. MIT Lincoln Lab GMM toolkit is used for the UBM-GMM training. An independent set of 500 test trials is used to evaluate the verification performance. The number of impostor and genuine trials in the test set are 169 and 331 respectively. The data represents training and testing from an interview setting using the same microphone [66][6]. This condition is specifically chosen in order to focus on additive noise distortions, without introducing other channel mismatch scenarios in the standard NIST SRE - hence ensuring consistency *across* ASR and ASV results in noise. Also, the UBM-GMM recognition backend does not include factor analysis techniques [46] which address various channel mismatch scenarios present in the NIST SREs. Notice however that the UBM-GMM system used even without the factor analysis techniques achieves state-of-the-art recognition performance on the same microphone matched channel condition evaluated in this work.

2.4.3 Features

(i) For phoneme recognition experiments, each MFCC feature vector is obtained by stacking a set of 9 frames of standard 13 Mel frequency cepstral coefficients along with their first, second, and third order temporal derivatives[7]. The feature vector for the proposed system is obtained by taking the original 128 dimensions (32 auditory frequency channels x 4 scales, as described in Section 2.3 along with their first, second, and third order temporal derivatives.

(ii) For speaker recognition experiments, each MFCC feature vector is obtained by

[6] corresponds to condition 2 of the eight common conditions evaluated in the NIST 2008 speaker recognition evaluation

[7] the 9 frame context window and the resulting 468 dimensional feature representation achieved best recognition performance, better than the standard 39 dimensional MFCC features [67]

taking 19 Mel frequency cepstral coefficients along with their first and second order temporal derivatives. Note that the higher order cepstral coefficients are more common in the speaker recognition literature and form the state-of-the-art feature representation in recent NIST SREs. Similarly, the feature vector for the proposed system is obtained by taking the base feature representation along with its first and second order temporal derivatives.

2.5 Recognition Results

2.5.1 Performance of multi-resolution features

Extending the mutual information analysis presented in the Section 2.3.4, we empirically show the relevance of set of scales $\{0.25, 0.5, 1.0, 2.0\}CPOand\{0.5, 1.0, 2.0, 4.0\}$ CPO for speech and speaker recognition tasks respectively. The performance of the multi-resolution features that encode these two sets of scales for the ASR and ASV tasks is shown in Table 2.1. Notice in particular how encoding the lower scales and omitting the higher scales improved the speech recognition performance, and vice-versa for speaker recognition task.

Table 2.1: Automatic Speech Recognition (ASR) and and Automatic Speaker Verification (ASV) performance of the proposed multi-resolution features. ASR performance is shown in Phoneme Recognition Rate (PRR) and ASV performance is shown in Equal Error Rate (EER).

Scales encoded	ASR Performance	ASV Performance
in the features (CPO)	(in PRR, %)	(in EER, %)
$[0.25, 0.5, 1, 2]$	**71.9**	3.4
$[0.5, 1, 2, 4]$	68.7	**2.7**

2.5.2 Comparison with standard front-end features

The proposed multi-resolution features are contrasted with MFCC features on both ASR and ASV tasks. To evaluate the noise robustness aspect of the two feature representations, various noisy versions of the test set are created by adding four types of noises at Signal-to-Noise-Ratio (SNR) levels of 20dB, 15dB, and 10dB. The noise types chosen are, Factory floor noise (Factory1), Speech babble noise (Babble), Volvo car interior noise (Volvo), and F16 cockpit noise (F16), all taken from NOISEX-92 database, and added using the standard FaNT tool [68]. In all the experiments, the recognition models are trained only on the original clean training set and tested on the clean as well as noisy versions of test set (*mismatch* train and test conditions). The phoneme recognition accuracy and speaker verification performance of the MFCCs and the multi-resolution features is listed in Table 2.2. The proposed multi-resolution features achieve ASR and ASV performance comparable to that of MFCCs under clean conditions. With additive noise conditions reflecting a variety of real acoustic scenarios, the multi-resolution features perform substantially better than the MFCCs - an average relative improvement of 38.9% on the ASR task and an average relative error rate reduction of 31.9% on the ASV task.

2.5.3 Comparison with state-of-the-art noise robust scheme

We further compare the performance of multi-resolution features with a state-of-the-art noise robust feature scheme, Mean-Variance ARMA (MVA) processing of MFCC features [20]. The MVA processing, when applied with the standard MFCC features, combines the advantages of multiple noise robustness schemes: cepstral mean subtraction, variance normalization, and temporal modulation filtering. The MVA has been shown to provide excellent robustness for additive noise distortions and form the state-of-the-art in noise robustness evaluations on the Aurora 2.0 and Aurora 3.0 databases [20]. Note that the auto-regression-moving-average (ARMA) filtering in the MVA processing is shown to be superior to temporal modulation filtering techniques like RASTA [69] for noise robustness.

To further improve the noise robustness of multi-resolution features and be consistent with the temporal modulation filtering employed in the MVA feature scheme, the multi-resolution features are processed with a bandpass modulation filter applied in the temporal domain [8]. The filtering is done in the Fourier domain of the mod-

[8] Note that the MVA processing has been shown to be *optimal* for cepstral domain features [20].

Table 2.2: Automatic Speech Recognition (ASR) and and Automatic Speaker Verification (ASV) performance of MFCC and multi-resolution feature representations for different types of noise.

Noise	SNR (dB)	ASR Performance (PRR)		ASV Performance (EER)	
		MFCC	Proposed	MFCC	Proposed
Clean	∞	71.4	71.9	2.7	2.7
Factory1	20	48.2	61	7.1	5.9
	15	38.1	53.1	10.9	7.6
	10	28.3	42.7	17.8	11.4
	5	19.6	30.9	28.4	18.7
	Average	33.5	46.9	16.1	10.9
Babble	20	48.1	64.1	5.4	4.1
	15	37.3	55.8	7.9	5.9
	10	27.6	43.7	11.5	9.7
	5	19.5	29	24.8	14.2
	Average	33.1	48.1	12.4	8.4
Volvo	20	60.8	70.9	3.9	2.9
	15	55.7	70.7	4.6	3.4
	10	49.9	70.1	6.4	4.8
	5	42.9	68.9	10.9	6.5
	Average	52.3	70.1	6.4	4.4
F16	20	48.5	61.4	10.7	7.5
	15	37.8	53.3	16.3	10.7
	10	27	40.9	21.7	14.5
	5	18.2	27.2	29.9	21.1
	Average	32.8	45.7	19.6	13.4

ulation amplitude. First the Fourier transform of the time sequence of each feature in the feature stream is taken, then is multiplied by a bandpass modulation filter $H_T(w;[0.5,12])$ capturing the modulation content within the specified range of 0.5Hz and 12Hz. Note that this temporal modulation range has been shown to be *information rich* and crucial for speech comprehension [56]. The inverse Fourier transform then yields the modulation filtered version of the feature stream. The bandpass modulation filter $H_T(w;[0.5,12])$ is defined as follows:

$$\begin{aligned} H_T(w;[0.5,12]) &= (\alpha w)^2\, e^{\left[1-(\alpha w)^2\right]}, \qquad (2.4) \\ \alpha &= \begin{cases} 1/0.5, & 0 \leq w < 0.5 \\ 1/w, & 0.5 \leq w \leq 12 \\ 1/12, & 12 < w \leq w_{max}, \end{cases} \end{aligned}$$

where w_{max} is the modulation frequency resolution - 50Hz corresponding to the 10ms frame-rate of the feature stream.

The phoneme recognition accuracy and speaker verification performance of MVA and *enhanced* multi-resolution features (E_MRF) is shown in Table 2.3. In addition to being comparable in the clean/matched conditions, the E_MRF features perform significantly better than MVA features in noisy/mismatch conditions - an average relative improvement of 12.2% on the ASR task and an average relative error rate reduction of 33.9% on the ASV task.

Consequently, it may be sub-optimal to apply the same processing on the multi-resolution features.

Table 2.3: Automatic Speech Recognition (ASR) and and Automatic Speaker Verification (ASV) performance of MFCC_MVA and E_MRF representations for different types of noise.

Noise Type	SNR (dB)	ASR Performance (PRR)		ASV Performance (EER)	
		MFCC_MVA	E_MRF	MFCC_MVA	E_MRF
Clean	∞	68.2	69.5	3	2.9
Factory1	20	55.7	61.7	5.4	5.2
	15	48.4	55.3	10	6.5
	10	39.4	45.5	16.6	10.7
	5	30.2	34.3	23.9	16.3
	Average	43.4	49.2	13.9	9.6
Babble	20	56.5	64.5	4.5	3.9
	15	49.5	57.7	6.2	5.4
	10	40.7	48.1	10.7	8.9
	5	29.7	34.4	19.5	12.4
	Average	44.1	51.1	10.2	7.6
Volvo	20	63.5	69.4	3.6	3
	15	62	69.2	5.2	3.4
	10	60.2	68.6	6.5	4.6
	5	58.1	67.7	9.4	6.2
	Average	60.9	68.7	6.1	4.3
F16	20	57.1	61.8	12.4	7.3
	15	50.8	55.6	18.3	10.2
	10	43.2	46.4	22.4	12.4
	5	34.6	35.1	26.6	16.6
	Average	46.4	49.7	19.9	11.6

2.6 Discussion

In this work, we begin to address the issue of versatile speech representations that could bear relevance to both speaker and speech recognition tasks. The proposed scheme captures the prominent features of the speech spectrum ranging from its broad trends (which correlate with vocal tract shape and length) to its rapidly varying details (which capture information about harmonics and voice quality). Because of the non-targeted nature of the proposed multi-resolution analysis, it is able to map the speech signal onto a rich space that highlights information about the glottal shape and movements as well as vocal tract geometry and articulatory configuration. Notice how the proposed analysis allowed for defining two slightly different feature representations for speech and speaker recognition tasks using the same feature analysis machinery. This multi-resolution representation can be viewed as a *local* variant (w.r.t log-frequency axis) of the analysis provided by the cepstral decomposition (MFCC). Spectral shape information in cepstral analysis is scattered over all cepstrum coefficients and hence must be considered collectively, and not individually. In the proposed localized approach, one can mine the information in each scale component individually. While the two methods perform comparably in clean, the proposed feature representations reveal substantial robustness under noisy conditions in both ASR and ASV tasks.

The current effort is not the first attempt at bringing more biological realism to analysis of speech signals. A number of authors have explored improvements to speech feature analysis that ranged from detailed modeling of the efferent auditory periphery, including intricate nonlinear effects and firing patterns at the auditory nerve [70–74], cochleogram-type representations [75], stabilized and normalized auditory image representations [76], to even more selective model-based spectro-temporal fragments and dynamic maps [77, 78]. Auditory-inspired techniques have generally led to noticeable improvements over more 'conventional' signal processing methods for recognition tasks, particularly when dealing with distorted signals in presence of background or competing noises [16, 79, 80]. Additional techniques have also been proposed to take advantage of the multi-resolution scheme taking place at more central stations of the auditory pathway; whereby the spectral details of the signal as they evolve over time are meticulously analyzed via parallel channels that capture intricate details of the signal of interest. Recent implementations of such schemes have been shown to yield noticeable improvements to automatic speech recognition, particularly with regards to its noise-robustness [49]. The current work falls in the same category of more centrally-inspired analysis of speech signals. It provides two major advantages over comparable methods [49,50]: It does not involve dimension-expanded representations (close to 30,000 dimensions) which would inherently require tedious and computationally-expensive schemes hence limiting their applicability. Instead, our model is constrained to a perceptually-relevant spectral modulation subspace and further uses a judicious sampling scheme to encode the information with only four modulation filters. This results in a low-dimensional and highly robust feature space. The enhanced multi-resolution features also constrain temporal modulations to a perceptually-relevant space shown to be crucial for speech comprehension. Note that none of the components of the model have been calibrated to deal with a specific noise condition making it appropriate for testing in a wide range of acoustic environments.

Our ongoing efforts are aimed at achieving further improvements by applying the multi-resolution analysis on enhanced spectral profiles obtained from speech enhancement techniques [22] that benefit from additional voice/speech activity detectors and noise estimation/compensation techniques. Also, the noise robustness obtained here from the proposed multi-resolution features can extend to other large scale ASR tasks in TANDEM framework [81]. Similarly, more elaborate ASV systems are achievable using multi-resolution features in conjunction with standard practices in speaker recognition like factor analysis, supervectors and score normalization [46].

Chapter 3

Multi-dimensional Representations for Music Sound Objects

3.1 Chapter Outline

Timbre is the attribute of sound that allows humans and other animals to distinguish among different sound sources. Studies based on psychophysical judgments of musical timbre, ecological analyses of sound's physical characteristics as well as machine learning approaches have all suggested that timbre is a multifaceted attribute that invokes both spectral and temporal sound features. Here, we explored the neural underpinnings of musical timbre. We used a neuro-computational framework based on spectro-temporal receptive fields, recorded from over a thousand neurons in the mammalian primary auditory cortex as well as from simulated cortical neurons, augmented with a nonlinear classifier. The model was able to perform robust instrument classification irrespective of pitch and playing style, with an accuracy of 98.7%. Using the same front end, the model was also able to reproduce perceptual distance judgments between timbres as perceived by human listeners. We also extended this model to continous musical performances and proposed a note extraction technique along with model adaptation to be able to capture the musical instrument identitiy. The study demonstrates that joint spectro-temporal features, such as those observed in the mammalian primary auditory cortex, are critical to provide the rich-enough representation necessary to account for perceptual judgments of timbre by human listeners, as well as recognition of musical instruments.

3.2 Introduction

A fundamental role of auditory perception is to infer the likely source of a sound; for instance to identify an animal in a dark forest, or to recognize a familiar voice on the phone. Timbre, often referred to as the color of sound, is believed to play a key role in this recognition process [82]. Though timbre is an intuitive concept, its formal definition is less so. The ANSI definition of timbre describes it as that attribute that allows us to distinguish between sounds having the same perceptual duration, loudness, and pitch, such as two different musical instruments playing exactly the same note [83]. In other words, it is neither duration, nor loudness, nor pitch; but is likely "everything else".

As has been often been pointed out, this definition by the negative does not state what are the perceptual dimensions underlying timbre perception. Spectrum is obviously a strong candidate: physical objects produce sounds with a spectral profile that reflects their particular sets of vibration modes and resonances [84]. Measures of spectral shape have thus been proposed as basic dimensions of timbre (e.g., formant position for voiced sounds in speech, sharpness, and brightness) [2, 3]. But timbre is not only spectrum, as changes of amplitude over time, the so-called temporal envelope, also have strong perceptual effects [4, 5]. To identify the most salient timbre dimensions, statistical techniques such as multidimensional scaling have been used: perceptual differences between sound samples were collected and the underlying dimensionality of the timbre space inferred [6, 7]. These studies suggest a combination of spectral and temporal dimensions to explain the perceptual distance judgments, but the precise nature of these dimensions varies across studies and sound sets [9, 85]. Importantly, almost all timbre dimensions that have been proposed to date on the basis of psychophysical studies [86] are either purely spectral or purely temporal. The only spectro-temporal aspect of sound that has been considered in this context is related to the asynchrony of partials around the onset of a sound [6, 7], but the salience of this spectro-temporal dimension was found to be weak and context-dependent [87].

Technological approaches, not concerned with biology nor human perception, have explored much richer feature representations that span both spectral, temporal, and spectro-temporal dimensions. The motivation for these engineering techniques is an accurate recognition of specific sounds or acoustic events in a variety of applications (e.g. automatic speech recognition; voice detection; music information retrieval; target tracking in multisensor networks and surveillance systems; medical diagnosis, etc.). Myriad spectral features have been proposed for audio content analysis, ranging from simple summary statistics of spectral shape (e.g. spectral amplitude, peak, centroid, flatness) to more elaborate descriptions of spectral information such as Mel-Frequency Cepstral Coefficients (MFCC) and Linear or Perceptual Predictive Coding (LPC or PLP) [88–90]. Such metrics have often been augmented with temporal information, which was found to improve the robustness of content identification [91, 92]. Common modeling of temporal dynamics also ranged from simple summary statistics such as onsets, attack time, velocity, acceleration and higher-order moments to more sophisticated statistical temporal modeling using Hidden Markov Models, Artificial Neural Networks, Adaptive Resonance Theory models, Liquid State Machine systems and Self-Organizing Maps [93, 94]. Overall, the choice of features was very dependent on the task at hand, the complexity of the dataset, and the desired performance level and robustness of the system.

Complementing perceptual and technological approaches, brain-imaging techniques have been used to explore the neural underpinnings of timbre perception. Correlates of musical timbre dimensions suggested by multidimensional scaling studies have been observed using event-related potentials [10]. Other studies have attempted to identify the neural substrates of natural sound recognition, by looking for brain areas that would be selective to specific sound categories, such as voice-specific regions in secondary cortical areas [11, 12] and other sound categories such as tools [13] or musical instruments [14]. A hierarchical model consistent with these findings has been proposed in which selectivity to different sound categories is refined as one climbs the processing chain [95]. An alternative, more distributed scheme has also been suggested [96,97], which includes the contribution of low-level cues to the large perceptual differences between these high-level sound categories.

A common issue for the psychophysical, technological, and neurophysiological investigations of timbre is that the generality of the results is mitigated by the particular characteristics of the sound set used. For multi-dimensional scaling behavioral studies, by construction, the dimensions found will be the most salient within the sound set; but they may not capture other dimensions which could nevertheless be crucial for the recognition of sounds outside the set. For engineering studies, dimensions may be designed arbitrarily as long as they afford good performance in a specific task. For the imaging studies, there is no suggestion yet as to which low-level acoustic features may be used to construct the various selectivity for high-level categories while preserving invariance within a category. Furthermore, there is a major gap between these studies and what is known from electrophysiological recordings in animal models. Decades of work have established that auditory cortical responses display rich and complex spectro-temporal receptive fields, even within primary areas [30, 98]. This seems at odds with the limited set of spectral or temporal dimensions that are classically used to characterize timbre in perceptual studies.

To bridge this gap, we investigate how cortical processing of spectro-temporal modulations can subserve both sound source recognition of musical instruments and perceptual timbre judgments. Specifically, cortical receptive fields and computational models derived from them are shown to be suited to classify a sound source from its evoked neural activity, across a wide range of instruments, pitches and playing styles, and also to predict accurately human judgments of timbre similarities

3.3 Methods

3.3.1 Psychoacoustics

3.3.1.1 Stimuli

Recordings of single musical notes were extracted from the RWC Music Database [99], using the notes designated "medium-volume" and "staccato", with pitches of A3, D4, and G#4. The set used for the experiments comprised 13 sound sources: Piano, Vibraphone, Marimba, Cello, Violin, Oboe, Clarinet, Trumpet, Bassoon, Trombone, Saxophone, male singer singing the vowel /a/, male singer singing the vowel /i/. Each note was edited into a separate sound file, truncated to 250 ms duration with 50ms raised cosine offset ramp (the onset was preserved), and normalized in RMS power.

More details on the sound set can be found in [100]. The analyses presented in the current study exclude the results from the 2 vowels, as only musical instruments were considered in the model classification experiments.

3.3.1.2 Participants and Apparatus

A total of twenty listeners participated in the study (14 totally nave participants, 6 participants experienced in psychoacoustics experiment but nave to the aim of the present study; mean age: 28y; 10 female). They had no self-reported history of hearing problems. All twenty subjects performed the test with the D4 pitch. Only six took part in the remaining tests with notes A3 and G#4. Results from the 6 subjects tested on all 3 notes are reported here, even though we checked that including all subjects would not change our conclusions. Stimuli were played through an RME Fireface sound-card at a 16-bit resolution and a 44.1 kHz sample-rate. They were presented to both ears simultaneously through Sennheiser HD 250 Linear II headphones. Presentation level was 65 dB(A). Listeners were tested individually in a double-walled IAC sound booth.

3.3.1.3 Procedure

Subjective similarity ratings were collected. For a given trial, two sounds were played with a 500 ms silent interval. Participants had to indicate how similar they perceived the sounds to be. Responses were collected by means of a graphical interface with a continuous slider representing the perceptual similarity scale. The starting position of the slider was randomized for each trial. Participants could repeat the sound pair as often as needed before recording their rating. In an experimental block, each sound was compared to all others (with both orders of presentations) but not with itself. This gave a total of 156 trials per block, presented in random order. Before collecting the experimental data, participants could hear the whole sound set three times. A few practice trials were also provided until participants reported having understood the task and instructions. A single pitch was used for all instruments in each block; the three types of blocks (pitch A3, D4, or G#4) were run in counterbalanced order across participants. Two blocks per pitch were run for each participants, and only the second block was retained for the analysis.

3.3.1.4 Multidimensional scaling (MDS) and acoustical correlates

To compare the results with previous studies, we ran an MDS analysis on the dissimilarity matrix obtained from human judgments. A standard non-metric MDS was performed (Matlab, the MathWorks). Stress values were generally small, with a knee-point for the solution at two dimensions (0.081, Kruskal normalized stress1). We also computed acoustical descriptors corresponding to the classic timbre dimensions. Attack time was computed by taking the logarithm of the time taken to go from -40dB to -12dB relative to the maximum waveform amplitude. Spectral centroid was computed by running the stimuli in an auditory filterbank, compressing the energy in each channel (exponent: 0.3), and taking the center of mass of the resulting spectral distribution.

3.3.2 Auditory Model

The cortical model is comprised of two main stages: an early stage mimicking peripheral processing up to the level of the midbrain, and a central stage capturing processing in primary auditory cortex (A1). This model is described in Section 1.2 and is given by Equations 1.1,1.2. The present study uses 11 spectral filters with characteristic scales [0.25, 0.35, 0.50, 0.71, 1.00, 1.41, 2.00, 2.83, 4.00, 5.66, 8.00] (cycles/octave) and 11 temporal filters with characteristic rates [4.0, 5.7, 8.0, 11.3, 16.0, 22.6, 32.0, 45.3, 64.0, 90.5, 128.0] (Hz), each with upward and downward directionality. All outputs are integrated over the time duration of each note. In order to simplify the analysis, we limit our computations to the magnitude of the cortical output $r_{\pm}(t, f; \omega, \Omega; \theta, \phi)$ (i.e. responses corresponding to zero-phase filters).

Finally, dimensionality reduction is performed using tensor singular-value decomposition [101]. This technique unfolds the cortical tensor along each dimension (frequency, rate and scale axes) and applies singular value decomposition on the unfolded matrix. We choose 5 eigenscales, 4 eigenrates and 21 eignefrequencies resulting in 420 features with the highest eigenvalues, preserving 99.9% of the variance in the original data. The motivation for this cutoff choice is presented later.

3.3.3 Cortical receptive Fields

Data used here was collected in the context of a number of studies [37, 102, 103] and full details of the experimental paradigm are described in these publications. Briefly, extracellular recordings were performed in 15 awake non-behaving domestic ferrets (Mustela putorius) with surgically implanted headposts. Tungsten electrodes (3-8 M) were used to record neural responses from single and multi-units at different depths. All data was processed off-line and sorted to extract single-unit activity.

Spectro-Temporal Receptive fields (STRF) were characterized using TORC (Temporally-Orthogonal Ripple Combination) stimuli [104], consisting of superimposed ripple noises with rates between 4-24 (Hz) and scales between 0 (flat) and 1.4 peaks/octave. Each stimulus was 3 sec with inter-stimulus intervals of 1 -1.2sec, and a full set of 30 TORCs was typically repeated 6-15 times. All sounds were computer-generated and delivered to the animal's ear through inserted earphones calibrated in-situ. TORC amplitude is fixed between 55-75 dB SPL.

STRFs were derived using standard reverse correlation techniques, and a signal-to-noise ratio (SNR) for each STRF was measured using a bootstrap technique (see [104] for details). Only STRFs with SNR$\geq$2 were included in the current study, resulting in a database of 1110 STRFs (average 74 STRFs/animal). Note because of the experimental paradigm, STRFs spanned a 5-octave range with low frequencies 125, 250 or 500Hz. In the current study, all STRFs were aligned to match the frequency range of musical note spectrograms. Since all our spectrograms start at 180Hz and cover 5.3 octaves, we scaled and shifted the STRF's to fit this range.

The neurophysiological STRFs were employed to perform the timbre analysis by convolving each note's auditory spectrogram $z(t, f)$ with each STRF in the database as shown in Equation 3.1.

$$r_n(t, f) = z(t, f) * STRF(t, f)(2) \tag{3.1}$$

The resulting firing rate vector $r_n(t, f)$ was then integrated over time yielding an average response across the tonotopic axis. The output from all STRFs were then stacked together, resulting in a 142080 (128 frequency channels x1110 STRFs) dimensional vector. We reduced this vector using singular value decomposition and mapped it onto 420 dimensions, which preserve 99.9% of the data variance in agreement with dimensionality used for model STRFs.

3.3.4 Timbre Classification

In order to test the cortical representation's ability to discriminate between different musical instruments, we augmented the basic auditory model with a statistical clustering model based on support vector machines (SVM) [105]. Support vector machines are classifiers that learn a set of hyperplanes (or decision boundaries) in order to maximally separate the patterns of cortical responses caused by the different instruments.

Each cortical pattern was projected via Gaussian kernel to a new dimensional space. The use of kernels is a standard technique used with support vector machines, aiming to map the data from its original space (where data may not be linearly separable) onto a new representational space that is linearly separable. This mapping of data to a new (more linear space) through a the use of a kernel or transform is commonly referred to as the "kernel trick" [105]. In essence, kernel functions aim to determine the relative position or similarity between pairs of points in the data. Because the data may lie in a space that is not linearly separable (not possible to use simple lines or planes to separate the different classes), it is desirable to map the data points onto a different space where this linear separability is possible. However, instead of simply projecting the data points themselves onto a high-dimensional feature space which would increase complexity as a function of dimensionality, the "kernel trick" avoids this direct mapping. Instead, it provides a method for mapping the data into an inner product space without explicitly computing the mapping of the observations directly. In other words, it computes the inner product between the data points in the new space without computing the mapping explicitly. The kernel used here is given by Equation 3.2

$$K(x, y) = \exp\left(\frac{-(x-y)^T(x-y)}{\sigma}\right) \tag{3.2}$$

where x and y are the feature vectors of 2 sound samples. The parameter for the Gaussian kernel and the cost parameter for the SVM algorithm were optimized on a subset of the training data. A classifier C_{ij} is trained for every pair of classes i and j. Each of these classifiers then gives a label l_{ij} for a test sample. Note that l_{ij}=i or j. We count the number of labels $R_i = \sum 1(l_{ij} = i)$. The test sample is then assigned to the class with maximum count given by $\arg\max_i(R_i)$. The parameter σ in Equation 3.2 was chosen by doing a grid search over a large parameter span in order to optimize the classifier performance in correctly distinguishing different instruments. This tuning was done by training and testing on a subset of the training data. For model testing, we performed a standard k-fold cross validation procedure with k=10 (90% training, 10% testing). The dataset was divided into 10 parts. We then left out one part at a time and trained on the remaining 9 parts. The results reported are the average performance over all 10 iterations. A single Gaussian parameter was optimized for all the pair-wise classifiers across all the 10-fold cross validation experiments.

3.3.5 Analysis of Support Vector distribution

In order to better understand the mapping of the different notes in the high-dimensional space used to classify them, we performed a closer analysis of the support vectors for each instrument pair i and j. Support vectors are the samples from each class that fall exactly on the margin between class i and class j, and therefore are likely to be more confusable between the classes. Since we are operating in the 'classifier space', each of the support vectors is defined in a reduced dimensional hyperspace consisting of 5 eigen-scales, 4 eigen-rates, and 21 eigen-frequencies as explained above (a total of 420 dimensions). The collection of all support vectors for each class i can be pulled together to estimate a high-dimensional probability density function. The density function estimate was derived using a histogram method by partitioning the sample space along each dimension into 100 bins, counting how many samples fall into each bin and dividing the counts by the total number of samples. We label the probability distribution for the d-th dimension (d = 1,..,420) $p_{i,d}$. We then computed the symmetric KL divergence, $KL_{i,j}(d)$ [58], between the support vectors for classes i and j from the classifier C_{ij} as shown in Equation 3.3. The KL divergence is simply a measure of difference between pairs of probability distributions as defined in Equation 3.3.

$$KL_{i,j}(d) = p_{i,d}log\left(\frac{p_{i,d}}{p_{j,d}}\right) + p_{j,d}log\left(\frac{p_{j,d}}{p_{i,d}}\right) \tag{3.3}$$

The bins with zero probability were disregarded from the computation of the KL divergence. An alternative method that smoothed the probability distribution over the zero bins was also tested and yielded virtually comparable results. Overall, this analysis is meant to inform about the layout of the timbre decision space. We analyzed the significance of the results between the broad timbre classes (winds, percussions and strings) by pooling individual comparisons between instruments within each group (See Figure 3.5).

3.3.6 Dataset

We used the RWC music database [99] for testing the model. 11 instruments were used for this task, which included string (violin, piano, cello), percussion (vibraphone, marimba) and wind instruments (saxophone, trumpet, trombone, clarinet, oboe, and bassoon). We extracted an average of 1980 notes per instrument ranging over different makes of the instruments, as well as a wide range of pitches and styles of playing (staccato, vibrato, etc.). The notes were on average 2.7 sec in duration but varied between 0.1-18 sec. The sampling frequency of the wave files in the database was 44.1 kHz. We performed preprocessing on the sound files by first down sampling to 16 kHz then filtering using a pre-emphasis filter (FIR filter with coefficients 1 and -0.97).

3.3.7 Human vs. Model Correlation

We tested the auditory model's ability to predict human listeners' judgment of musical timbre distances. Just like the timbre classification task, we used the cortical model augmented with Gaussian Kernels. In order to optimize the model to the test data, we employed a variation of the Gaussian kernel that performs an optimized feature embedding on every data dimension as defined in Equation 3.4.

$$K(x,y) = \exp\,(-\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma_i}) \tag{3.4}$$

where N is the number of dimensions of the features x and y. σ_i's are parameters for the kernel that need to be optimized. We define an objective function that optimizes the correlation between the human perceptual distances and the distances in the embedded space.

$$J = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(K(x_i,x_j)-\overline{K}\right)\left(D(i,j)-\overline{D}\right)}{\left(\frac{n(n-1)}{2}-1\right)\sigma_K\sigma_D} \tag{3.5}$$

where x_i is the average profile for the ith instrument over all notes; $D(i,j)$ is the average perceived distance between the ith and jth instrument based on psychoacoustic results $\overline{K}$ and $\overline{D}$ are the average distances from the kernel and the psychoacoustic experiment respectively. σ_K represents the variance of the kernel distances over all samples (all instrument pairs). Similarly σ_D is the variance of the human perceived distances. We used a gradient ascent algorithm to learn σ_i which optimize the objective function.

The correlation analysis employed the same dataset used for the human psychophysical experiment described above. Each note was 0.25s in duration with sampling rate 44.1 kHz and underwent the same preprocessing as mentioned earlier. The absolute value of the model output was derived for each note and averaged over duration following a similar procedure as the timbre classification described above. The cortical features obtained for the three notes (A3, D4, G#4) were averaged for each instrument i to obtain x_i. Similarly the perceived human distances between instrument i and j were obtained by averaging the (i,j)th and (j,i)th entry in the human distance matrix over all the 3 notes to obtain D(i,j).

Finally, the human and model similarity matrices were compared using the Pearson's correlation metric. In order to avoid overestimating the correlation between the two matrices (the two symmetric values appearing twice in the correlation), we correlated only the upper triangle of each matrix.

3.3.7.1 Dimensionality reduction of cortical features

As is the case with any classification problem in high-dimensional spaces, all analyses above had to be performed on a reduced number of features which we obtained using tensor singular value decomposition (TSVD), as described earlier. This step is necessary in order to avoid the curse of dimensionality which reduces the predictive power of the classifier as the dimensionality increases [106]. In order to determine the 'optimal' size of the reduced features, we ran a series of investigations with a range of TSVD thresholds. The analysis comparing the correlation between the cortical model and human judgments of timbre similarity is shown in Figure 3.8. The analysis led to the choice of 420 dimensions as near optimal. It is important to note that our tests were not fine-grained enough in order to determine the exact point of optimality. Moreover, this choice is only valid with regards to the data at hand and classifier used in this study, namely a support vector machine. If one were to choose a different classifier, the optimal reduced dimensionality may be different. It is merely a number that reflects the tradeoff between keeping a rich dimensionality that captures the diversity of the data; while reducing the dimensionality in order to fit the predictive power of the classifier.

To further emphasize this point, we ran a second analysis contrasting the system performance with the full cortical model (joint spectro-temporal modulations) against a model with separable modulations; all while maintaining the dimensionality of the reduced space fixed. This experiment (Figure 3.8 - red curve) confirmed that the original space indeed biases the system performance, irrespective of the size of reduced data. Results from Table 3.2 are also overlaid in the same figure for ease of comparison.

3.3.8 Control experiments

3.3.8.1 Auditory spectrum analysis

The auditory spectrum was obtained by analyzing the input waveform with the 128 cochlear filters described above, and integrating over the time dimension. The resulting feature vector was 128x1 representation of the spectral profile of each signal. Unlike a simple Fourier analysis of the signal, the cochlear filtering stage operated on a logarithmic axis with highly asymmetric filters.

3.3.8.2 Separable spectro-temporal modulation analysis

For an input spectrogram $y(t, f)$, the response of each rate filter (RF) and scale Filter(SF) was obtained separately as defined in Equations 3.6,3.7.

$$r_t(t, f; \omega; \theta) = y(t, f) *_t RF(t; \omega; \theta) \tag{3.6}$$

$$r_s(t, f; \Omega; \phi) = y(t, f) *_f SF(t; \Omega; \phi) \tag{3.7}$$

where $*_t$ denotes convolution in time, $*_f$ denotes convolution in frequency and θ is the characteristic phase of the RF and ϕ is the characteristic phase of the SF which determine the degree of asymmetry in the time and frequency axis respectively. Details of the design of the filter functions SF and RF can be found in [33]. Unlike the analysis given in Equation 1.2, the spectral and temporal modulations were derived separately using one-dimensional complex-valued filters (either along time or along frequency axis). The resulting magnitude outputs from Equations 3.7,3.6 were then stacked together to form the feature vector with 4224 (11scalesX128frequecies+22ratesX128frequecies) dimensions. The dimensionality was then reduced to 420 using tensor singular value decomposition retaining 99.9% of the variance.

3.3.8.3 Global modulation analysis

For this experiment, we used the $r_t(t, f; \omega; \theta)$ and $r_s(t, f; \Omega; \phi)$ from Equations 3.7,3.6, and integrated the output over time and frequency for each note. The resulting rate and scale responses were then stacked together to form the feature vector.

3.3.8.4 Under sampled joint modulation

In this experiment we aimed to make the dimensionality of the cortical model comparable to the separable model by under sampling the rate, scale and frequency axes. The auditory spectrogram was down sampled along the frequency axis by a factor of 2. This auditory spectrogram representation was then analyzed by 6 spectral filters with characteristic scales [0.25, 0.47, 0.87, 1.62, 3.03, 5.67] (cycles/octave) and 5 temporal filters with characteristic rates [4.0, 8.0, 16.0, 32.0, 64.0] (Hz), each with upward and downward directionality resulting in a 3840 dimensional representation. The dimensionality was then reduced to 420 using tensor singular value decomposition retaining 99.99% of the variance

3.4 Results

3.4.1 Cortical processing of complex musical sounds

Responses in primary auditory cortex (A1) exhibit rich selectivity that extends beyond the tonotopy observed in the auditory nerve. A1 neurons are not only tuned to the spectral energy at a given frequency, but also to the specifics of the local spectral shape such as its bandwidth [107], spectral symmetry [55], and temporal dynamics [108] (Figure 3.1). Put together, one can view the resulting representation of sound in A1 as a multidimensional mapping that spans at least three dimensions: (1) Best frequencies that span the entire auditory range; (2) Spectral shapes (including bandwidth and symmetry) that span a wide range from very broad (2-3 octaves) to narrowly tuned (¡ 0.25 octaves); and (3) Dynamics that range from very slow to relatively fast (1-30 Hz).

This rich cortical mapping may reflect an elegant strategy for extracting acous-

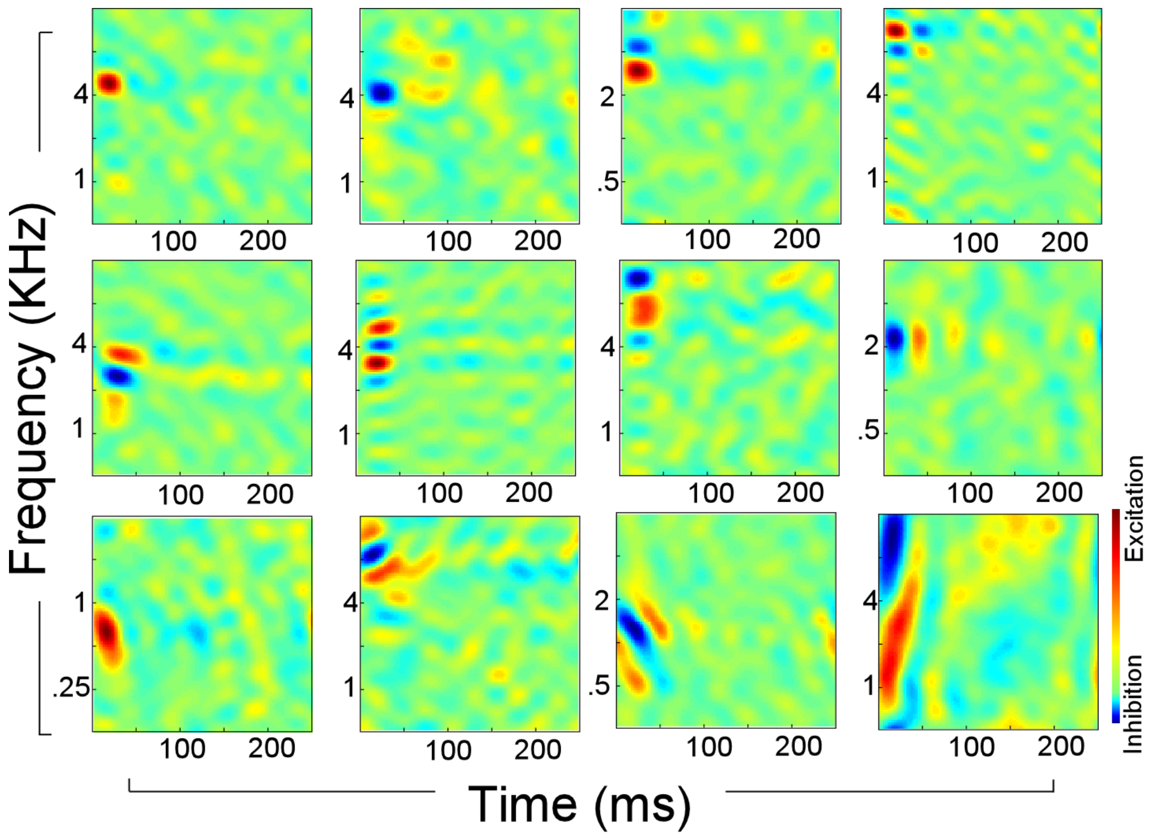

Figure 3.1: Neurophysiological receptive fields. Each panel shows the receptive field of 1 neuron with red indicating excitatory (preferred) responses, and blue indicating inhibitory (suppressed) responses. Examples vary from narrowly tuned neurons (top row) to broadly tuned ones (middle and bottom row). They also highlight variability in temporal dynamics and orientation (upward or downward sweeps).

tic cues that subserve the perception of various acoustic attributes (pitch, loudness, location, and timbre) as well as the recognition of complex sound objects, such as different musical instruments. This hypothesis was tested here by employing a database of spectro-temporal receptive fields (STRFs) recorded from 1110 single units in primary auditory cortex of 15 awake non-behaving ferrets. These receptive fields are linear descriptors of the selectivity of each cortical neuron to the spectral and temporal modulations evident in the cochlear ”spectrogram-like” representation of complex acoustic signals that emerges in the auditory periphery. Such STRFs (with a variety of nonlinear refinements) have been shown to capture and predict well cortical responses to a variety of complex sounds like speech, music, and modulated noise [109–113].

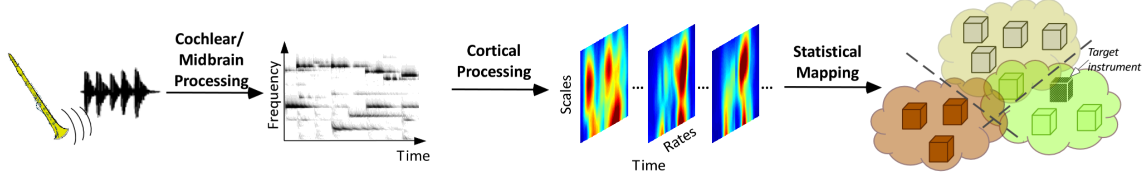

Figure 3.2: Schematic of the timbre recognition model. An acoustic waveform from a test instrument is processed through a model of cochlear and midbrain processing; yielding a time-frequency representation called auditory spectrogram. This later is further processed through the cortical processing stage through neurophysiological or model spectro-temporal receptive fields. Cortical responses of the target instrument are tested against boundaries of a statistical SVM timbre model in order to identify the instrument's identity.

To test the efficacy of STRFs in generating a representation of sound that can distinguish among a variety of complex categories, sounds from a large database of musical instruments were mapped onto cortical responses using the physiological STRFs described above. The time-frequency spectrogram for each note was convolved with each STRF in our neurophysiology database to yield a firing rate that is then integrated over time. This initial mapping was then reduced in dimensionality using singular value decomposition to a compact eigen-space; then augmented with a nonlinear statistical analysis using support vector machine (SVM) with Gaussian kernels [105] (see 3.3 for details). Briefly, support vector machines are classifiers that learn to separate, in our specific case, the patterns of cortical responses induced by the different instruments. The use of Gaussian kernels is a standard technique that allows

Model	Performance	
	Mean	STD
Auditory Spectrum (Gaussian kernel SVM)	79.1%	0.7%
Neurophysiological STRFs (Gaussian kernel SVM)	87.2%	0.8%
Full Cortical Model (Linear SVM)	96.2%	0.5%
Full Cortical Model (Gaussian kernel SVM)	98.7%	0.2%

Table 3.1: Classification performance for the different models. The middle column indicates the mean of the accuracy scores for the 10 fold cross validation experiment and the right column indicates their standard deviation. Models differ either in their feature set (e.g. full cortical model versus auditory spectrogram) or in the classifier used (linear SVM versus Gaussian kernel SVM).

to map the data from its original space (where data may not be linearly separable) onto a new representational space that is linearly separable. Ultimately, the analysis constructed a set of hyperplanes that outline the boundaries between different instruments. The identity of a new sample was then defined based on its configuration in this expanded space relative to the set of learned hyperplanes (Figure 3.2).

Based on the configuration above and a 10% cross-validation technique, the model trained using the physiological cortical receptive fields achieved a classification accuracy of 87.22% 0.81 (the number following the mean accuracy represents standard deviation, see Table 3.1). Remarkably, this result was obtained with a large database of 11 instruments playing between 30 and 90 different pitches with 3 to 19 playing styles (depending on the instrument), 3 style dynamics (mezzo, forte and piano), and 3 manufacturers for each instrument (an average of 1980 notes/instrument). This high classification accuracy was a strong indicator that neural processing at the level of primary auditory cortex could not only provide a basis for distinguishing between different instruments, but also had a robust invariant representation of instruments over a wide range of pitches and playing styles.

3.4.2 The cortical model

Despite the encouraging results obtained using cortical receptive fields, the classification based on neurophysiological recordings was hampered by various shortcomings including recording noise and other experimental constraints. Also, the limited selection of receptive fields (being from ferrets) tended to under-represent parameter ranges relevant to humans such as lower frequencies, narrow bandwidths (limited to a maximum resolution of 1.2 octaves), and coarse sampling of STRF dynamics.

To circumvent these biases, we employed a model that mimics the basic transformations along the auditory pathway up to the level of A1. Effectively, the model mapped the one-dimensional acoustic waveform onto a multidimensional feature space. Importantly, the model allowed us to sample the cortical space more uniformly than

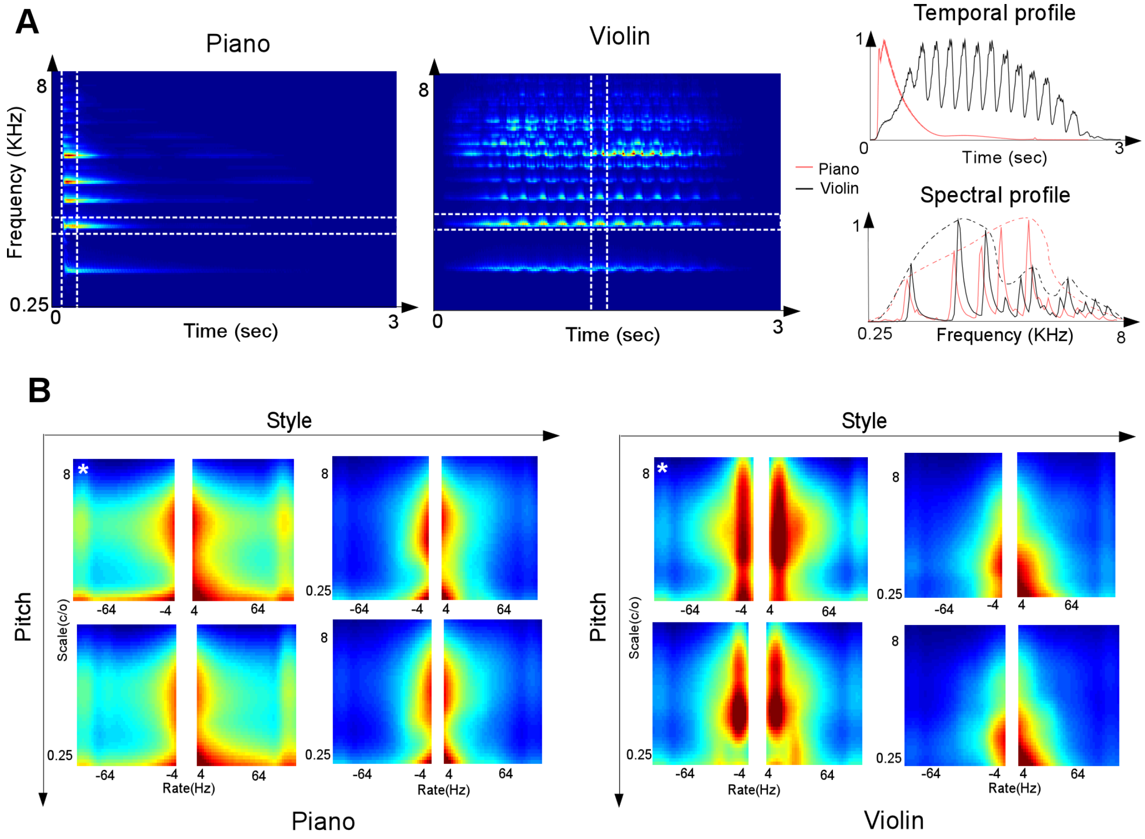

Figure 3.3: Spectro-temporal modulation profiles highlighting timbre differences between piano and violin notes. **(A)** The plot shows the time-frequency auditory spectrogram of piano and violin notes. The temporal and spectral slices shown on the right are marked. **(B)** The plots show magnitude cortical responses of four piano notes (left panels), played in normal (left) and Staccato (right) at F4 (top) and F#4 (bottom); and four violin notes (right panels), played in normal (left) and Pizzicatto (right) also at pitch F4(top) and F#4 (bottom). The white asterisks (upper leftmost notes in each quadruplet) indicate the notes shown in part **(A)** of this figure.

physiological data available to us, in line with findings in the literature [30, 98, 114].

The model operates by first mapping the acoustic signal into an auditory spectrogram. This initial transformation highlights the time varying spectral energies of different instruments which is at the core of most acoustic correlates and machine learning analyses of musical timbre [3, 9, 87, 115, 116]. For instance, temporal features in a musical note include fast dynamics that reflect the quality of the sound (scratchy, whispered, or purely voiced), as well as slower modulations that carry nuances of musical timbre such as attack and decay times, subtle fluctuations of pitch (vibrato) or amplitude (shimmer). Some of these characteristics can be readily seen in the auditory spectrograms, but many are only implicitly represented. For example, Figure 3.3A contrasts the auditory spectrogram of a piano vs. violin note. For violin, the temporal cross-section reflects the soft onset and sustained nature of bowing and typical vibrato fluctuations; the spectral slice captures the harmonic structure of the musical note with the overall envelope reflecting the resonances of the violin body. By contrast, the temporal and spectral modulations of a piano (playing the same note) are quite different. Temporally, the onset of piano rises and falls much faster, and its spectral envelope is much smoother.

The cortical stage of the auditory model further analyzes the spectral and temporal modulations of the spectrogram along multiple spectral and temporal resolutions. The model projects the auditory spectrogram onto a 4-dimensional space, representing time, tonotopic frequency, spectral modulations (or scales) and temporal modulations (or rates). The four dimensions of the cortical output can be interpreted in various ways. In one view, the cortical model output is a parallel repeated representation of the auditory spectrogram viewed at different resolutions. A different view is one of a bank of spectral and temporal modulation filters with different tuning (from narrowband to broadband spectrally, and slow to fast modulations temporally). In such view, the cortical representation is a display of spectro-temporal modulations of each channel as they evolve over time. Ultimately each filter acts as a model cortical neuron whose output reflects the tuning of that neuronal site. The model employed here had 30,976 filters (128freq x 22 rates x 11 scales), hence allowing us to obtain a full uniform coverage of the cortical space and bypassing the limitations of neurophysiological data. Note that we are not suggesting that 30K neurons are needed for timbre classification, as the feature space is reduced in further stages of the model (see below). We have not performed an analysis of the number of neurons needed for such task. Nonetheless, a large and uniform sampling of the space seemed desirable.

By collapsing the cortical display over frequency and averaging over time, one would obtain a two-dimensional display that preserves the "global" distribution of modulations over the remaining two dimensions of scale and rates. This "scale-rate" view is shown in Figure 3.3B for the same piano and violin notes in Figure 3.3A as well as others. Each instrument here is played at two distinct pitches with two different playing styles. The panels provide estimates of the overall distribution of spectro-temporal modulation of each sound. The left panel highlights the fact that the violin vibrato concentrates its peak energy near 6 Hz (across all pitches and styles); which matches the speed of pulsating pitch change caused by the rhythmic rate of 6 pulses per second chosen for the vibrato of this violin note. By contrast, the rapid onset of piano distributes its energy across a wider range of temporal modulations. Similarly, the unique pattern of peaks and valleys in spectral envelopes of each instrument produces a broad distribution along the spectral modulation axis, with the violin's sharper spectral peaks activating higher spectral modulations while the piano's smoother profile activates broad bandwidths. Each instrument, therefore, produces a correspondingly unique spectro-temporal activation pattern that could potentially be used to recognize it or distinguish it from others.

3.4.3 Musical timbre classification

Several computational models were compared in the same classification task analysis of the database of musical instruments as described earlier with real neurophysiological data. Results comparing all models are summarized in Table 3.1. For what we refer to as the full model, we used the 4-D cortical model. The analysis started with a linear mapping through the model receptive fields, followed by dimensionality reduction and statistical classification using support vector machines with re-optimized Gaussian kernels (see 3.3). Tests used a 10% cross-validation method. The cortical model yielded an excellent classification accuracy of $98.7\% \pm 0.2$.

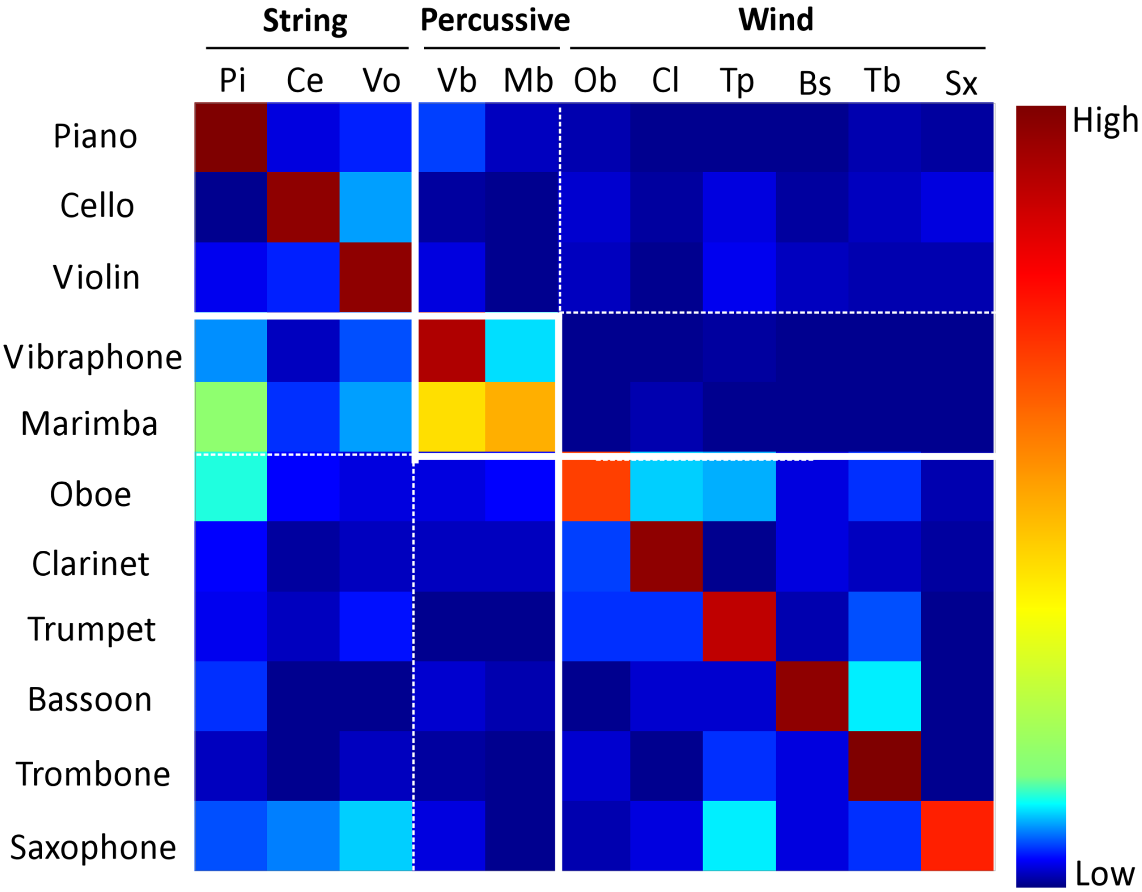

Figure 3.4: The confusion matrix for instrument classification using the auditory spectrum. Each row sums to 100% classification (with red representing high values and blue low values). Rows represent instruments to be identified and columns are instrument classes. Off diagonal values that are non-dark blue represent errors in classification. The overall accuracy from this confusion matrix is 79.1% 0.7.

We also explored the use of linear support vector machine, by bypassing the use of the Gaussian kernel. We performed a classification of instruments using the cortical responses obtained from the model receptive fields and a linear SVM. After optimization of the decision boundaries, we obtained an accuracy of 96.2% ± 0.5. This result supports our initial assessment that the cortical space does indeed capture most of the subtleties that are unique to a common instrument but distinct between different classes. It is mostly the richness of the representation that underlies the classification performance: only a small improvement in accuracy is observed by adding the non-linear warping in the full model.

In order to better understand the contribution of the cortical analysis beyond the time-frequency representation, we explored reduced versions of the full model. First we performed the timbre classification task using the auditory spectrogram as input. The feature spectra were obtained by processing the time waveform of each note through the cochlear-like filterbank front-end and averaging the auditory spectrograms over time, yielding a one-dimensional spectral profile for each note. These were then processed through the same statistical SVM model, with Gaussian functions optimized for this new representation using the exact same methods as used for cortical features. The classification accuracy for the spectral slices with SVM optimization attained a good but limited accuracy of 79.1%0.7. It is expected that a purely spectral model would not be able to classify all instruments. Whereas basic instrument classes differing by their physical characteristics (wind, percussion, strings) may have the potential to produce different spectral shapes, preserved in the spectral vector, more subtle differences in the temporal domain should prove difficult to recognize on this basis (see Figure 3.4). We shall revisit this issue of contribution and interactions between spectral and temporal features later (see 3.4.7).

We performed a post-hoc analysis of the decision space based on cortical features in an attempt to get a better understanding of the configuration of the decision hyperplanes between different instrument classes. The analysis treated the support vectors (i.e. samples of each instrument that fall right on the boundary that distinguishes it from another instrument) for each instrument as samples from an underlying high-dimensional probability density function. Then, a measure of similarity between pairs of probability functions (symmetric Kullback-Leibler (KL) divergence [58]) was employed to provide a sense of distance between each instrument pair in the decision space. Because of the size and variability in the timbre decision space, we pooled the comparisons by instrument class (winds, strings and percussions). We also focused our analysis on the reduced dimensions of the cortical space; called 'eigen'-rate, 'eigen'-scale and 'eigen'-frequencies; obtained by projecting the equivalent dimensions in the cortical tensor (rate, scale and frequency, respectively) into a reduced dimensional space using singular-value decomposition (see 3.3). The analysis revealed a number of observations (see Figure 3.5). For instance, wind and percussion classes were the most different (occupy distant regions in the decision space), followed by strings and percussions then strings and winds (average KL distances were 0.58, 0.41, 0.35, respectively). This observation was consistent with the subjective judgments of human listeners presented next (see off-diagonal entries in Figure 3.6B). All 3 pair comparisons were statistically significantly different from each other (Wilcoxon ranksum test, p¡10-5 for all 3 pairs). Secondly, the analysis revealed that the 2 first 'eigen'-rates captured most of the difference between the instrument classes (statistical significance in comparing the first 2 eigenrates with the others; Wilcoxon ranksum test, p=0.0046). In contrast, all 'eigen'-scales were variable across classes (Kruskal-Wallis test, p=0.9185 indicating that all 'eigen'-scales contributed equally in distinguishing the broad classes). A similar analysis indicated that the first four 'eigen'-frequencies were also significantly different from the remaining features (Wilcoxon ranksum test, p¡10-5). One way to interpret these observations is that the first two principal orientations along the rate axis captured most of the differences that distinguish winds, strings and percussions. This seems consistent with the large differences in temporal envelope shape for these instruments classes, which can be represented by a few rates. By contrast, the scale dimension (which captures mostly spectral shape, symmetry and bandwidth) was required in its entirety to draw a boundary between these classes, suggesting that unlike the coarser temporal characteristics, differentiating among instruments entails detailed spectral distinctions of a subtle nature.

3.4.4 Comparison with standard classification algorithms

Spectral features have been extensively used for tasks of musical timbre classification of isolated notes, solo performances or even multi-instrument recordings. Features such as Cepstral Coefficients or Linear Prediction of the spectrum resonances yielded performance in the range of 77% to 90% when applied to databases similar to the one used in the present study [117–119].

There is wide agreement in the literature that inclusion of simple temporal features, such as zero-crossing rate, or more complex ones such as trajectory estimation of spectral envelopes, is often desirable and results in improvement of the system performance. Tests on the RWC database with both spectral and temporal features reported an accuracy of 79.7% using 19 instruments [120] or 94.9% using 5 instruments [116]. Tests of spectrotemporal features on other music databases has often yielded a range of performances between 70-95% [121–124].

Whereas a detailed comparisons with our results is beyond the scope of this paper, we can still note that, if anything, the recognition rates we report for the full auditory model are generally in the range or above those reported by state-of-the-art signal processing techniques.

3.4.5 Psychophysics timbre judgments

Given the ability of the cortical model to capture the diversity of musical timbre across a wide range of instruments in a classification task, we next explored how well the cortical representation (from both real and model neurons) does in capturing human perceptual judgments of distance in the musical timbre space. To this end, we used human judgments of musical timbre distances using a psychoacoustic comparison paradigm.

Human listeners were asked to rate the similarity between musical instruments. We used three different notes (A3, D4 and G#4) in three different experiments. Similarity matrices for all three notes yielded reasonably balanced average ratings across subjects, instrument pair order (e.g. piano/violin vs. violin/piano) and pitches, in agreement with other studies [125] (Figure 3.6A). Therefore, we combined the matrices across notes and listeners into an upper half matrix shown in Figure 3.6B, and used it for all subsequent analyses. For comparison with previous studies, we also ran a multidimensional scaling (MDS) analysis [126] on this average timbre similarity rating and confirmed that the general configuration of the perceptual space was consistent with previous studies (Figure 3.6C) [6]. Also for comparison, we tested acoustical dimensions suggested in those studies. The first dimension of our space correlated strongly with the logarithm of attack-time (Pearson's correlation coefficient: =0.97, p¡10-3), and the second dimension correlated reasonably well with the center of mass of the auditory spectrogram, also known as spectral centroid (Pearson's correlation

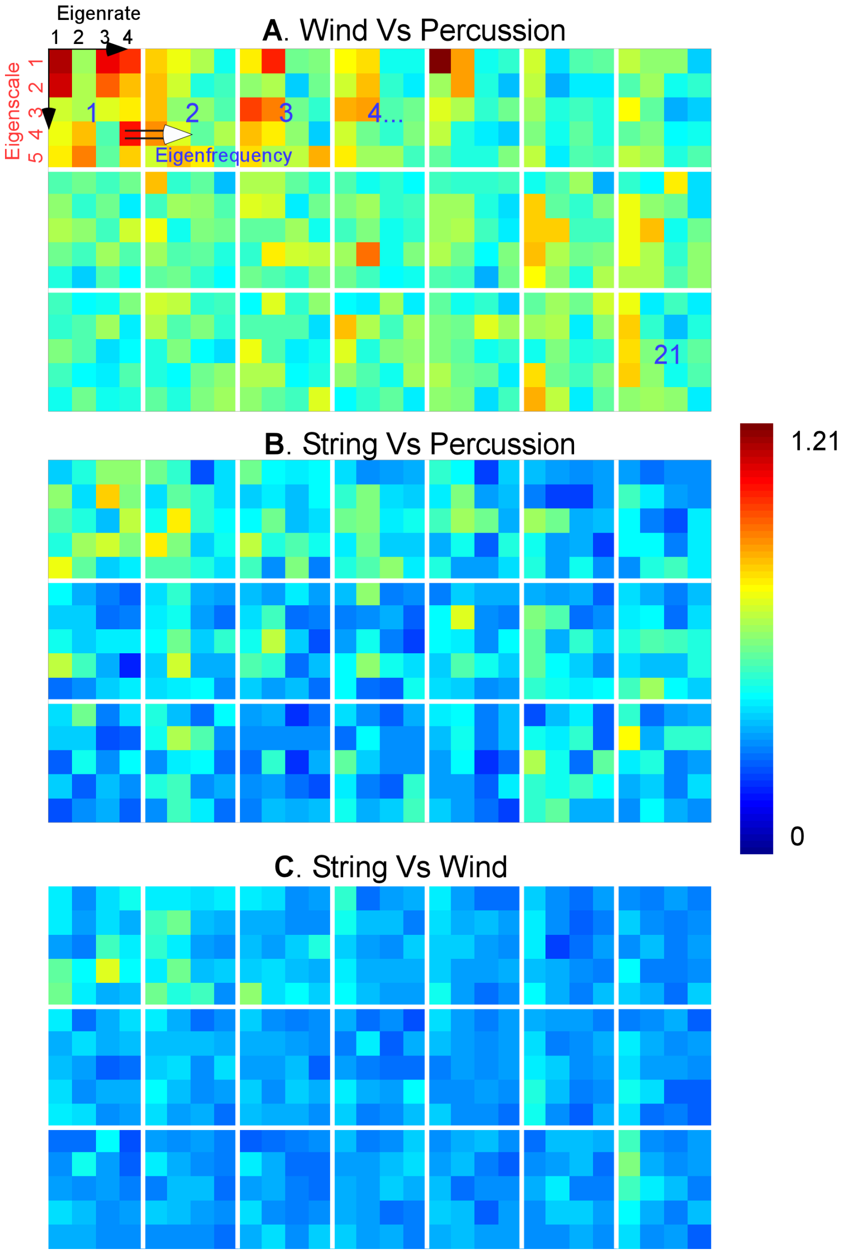

Figure 3.5: The average KL divergence between support vectors of instruments belonging to different broad classes. Each panel depicts the values of the 3 dimensional average distances between pairs of instruments of a given couple of classes: **(A)** wind vs. percussion; **(B)** string vs. percussion; **(C)** wind vs. string. The 3 dimensional vectors are displayed along eigenrates (x-axis), eigenscales (y-axis) and eigenfrequency (across small subpanels). Red indicates high values of KL divergence and blue indicates low values.

coefficient: =0.62, p=0.04).

3.4.6 Human vs. model timbre judgments

The perceptual results obtained above, reflecting subjective timbre distances between different instruments, summarizes an elaborate set of judgments that potentially reveal other facets of timbre perception than the listeners' ability to recognize instruments. We then explored whether the cortical representation could account for these judgments. Specifically, we asked whether the cortical analysis maps musical notes onto a feature space where instruments like violin and cello are distinct, yet closer to each other than a violin and a trumpet. We used the same 11 instruments and 3 pitches (A3, D4 and G#4) employed in the psychoacoustics experiment above and mapped them onto a cortical representation using both neurophysiological and model STRFs. Each note was then vectorized into a feature data-point and mapped via Gaussian kernels. These kernels are similar to the radial basis functions used in the previous section, and aimed at mapping the data from its original cortical space to a linearly separable space. Unlike the generic SVM used in the classification of musical timbre, the kernel parameters here were optimized based on the human scores following a similarity-based objective function. The task here was not merely to classify instruments into distinct classes, but rather to map the cortical features according to a complex set of rules. Using this learnt mapping, a confusion matrix was constructed based on the instrument distances, which was then compared with the human confusion matrix using a Pearson's correlation metric. We performed a comparison with the physiological as well as model STRFs. The simulated confusion matrices are shown in Figure 3.7A-B.

The success or otherwise of the different models was estimated by correlating the human dissimilarity matrix to that generated by the model. No attempt was made at producing MDS analyses of the model output, as meaningfully comparing MDS spaces is not a trivial problem [125]. Physiological STRFs yielded a correlation coefficient of 0.73, while model STRFs yielded a correlation of 0.94 (Table 3.2).

3.4.7 Control Experiments

In order to disentangle the contribution of the "input" cortical features versus the "back-end" machine learning in capturing human behavioral data, we recomputed confusion matrices using alternative representations such as the auditory spectrogram and various marginals of the cortical distributions. In all these control experiments, the Gaussian kernels were re-optimized separately to fit the data representation being explored.

We first investigated the performance using auditory spectrum features with optimized Gaussian kernels. The spectrogram representation yielded a similarity matrix that captures the main trends in human distance judgments, with a correlation coefficient of 0.74 (Figure 3.7C, leftmost panel). Similar experiments using a traditional spectrum (based on Fourier analysis of the signal) yield a correlation of 0.69.

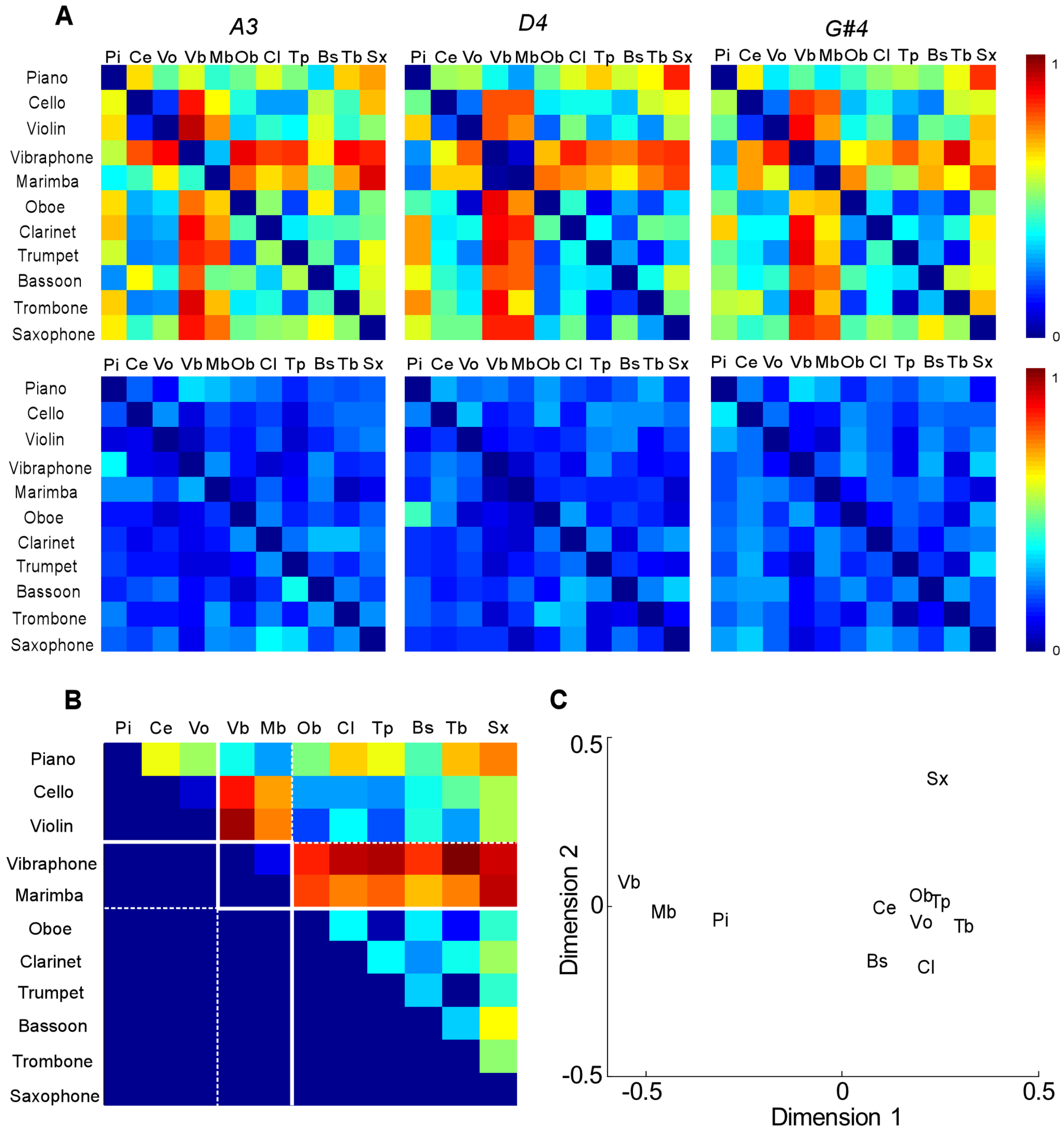

Figure 3.6: Human listener's judgment of musical timbre similarity. **(A)** The mean (top row) and standard deviation (bottom row) of the listeners' responses show the similarity between every pair of instruments for three notes A3, D4 and G#4. Red (values close to 1) indicates high dissimilarity and blue (values close to 0) indicates similarity. **(B)** Timbre similarity is averaged across subjects, musical notes and upper and lower half-matrices, and used for validation of the physiological and computational model. **(C)** Multidimensional scaling (MDS) applied to the human similarity matrix projected over 2 dimensions (shown to correlate with attack time and spectral centroid).

Feature Set	Distance Metric		
	L2 Features	L2 Reduced features	Gaussian kernel distance Reduced features
Fourier-based Spectrum	-	-	0.69
Auditory Spectrum	0.473	-	0.739
Global Spectro-temporal Modulation	0.509	-	0.701
Seperable Spectro-temporal Modulations	0.561	0.561	0.830
Full Cortical Model	0.611	0.607	0.944
Neurophysiological STRFs	-	-	0.73

Table 3.2: Correlation coefficients for different feature sets. Each row represents the correlation coefficient between the model and human similarity matrix using a direct Euclidian distance the specific feature itself (left); on a reduced dimension of the features (middle column) or using the Gaussian kernel distance (right column). Auditory Spectrum: time-average of the cochlear filterbank; Global Spectro-temporal Modulations: model STRFs averaged in time and frequency; Separable Spectro-temporal modulations: model STRFs averaged separately in rate and scale, and then in time; Full cortical model: STRFs averaged in time; Neurophysiological STRFs: as the full cortical model, but with STRFs collected in primary auditory cortex of ferrets.

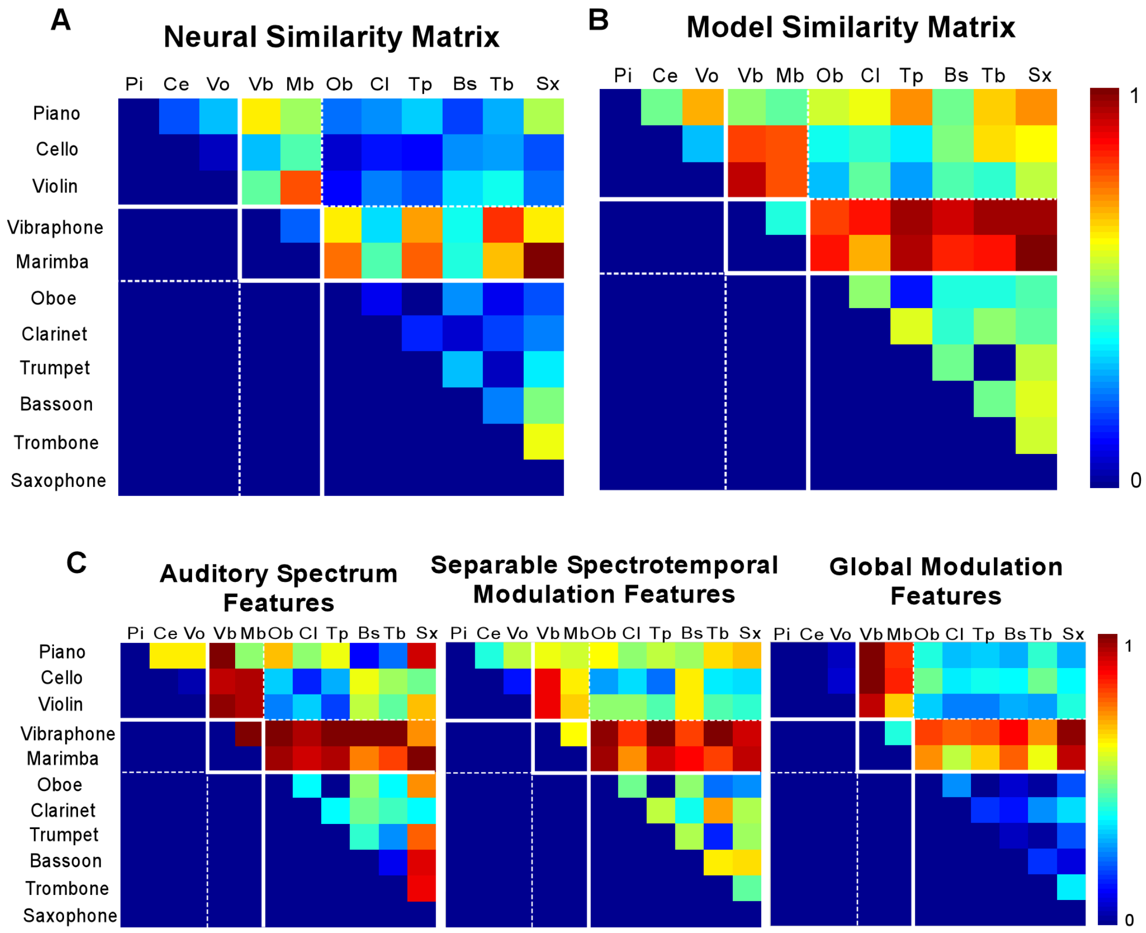

Figure 3.7: Model musical timbre similarity. Instrument similarity matrices based kernel optimization technique of the **(A)** neurophysiological receptive field and **(B)** cortical model receptive fields. **(C)** Control experiments using the auditory spectral features (left), separable spectro-temporal modulation feature (middle), and global modulation features [separable spectral and temporal modulations integrated across time and frequency] (right). Red depicts high dissimilarity. All the matrices show only the upper half-matrix with the diagonal not shown.

Next, we examined the effectiveness of the model cortical features by reducing them to various marginal versions with fewer dimensions as follows. First, we performed an analysis of the spectral and temporal modulations as a separable cascade of two operations. Specifically, we analyzed the spectral profile of the auditory spectrogram (scales) independently from the temporal dynamics (rates) and stack the two resulting feature vectors together. This analysis differed from the full cortical analysis that assumes an inseparable analysis of spectro-temporal features. An inseparable function is one that cannot be factorized into a function of time and a function of frequency; i.e. a matrix of rank greater than 1 (see 3.3). By construction, a separable function consists of temporal cross sections that are scaled versions of the same essential temporal function. A consequence of such constraint is that a separable function cannot capture orientation in time-frequency space (e.g. FM sweeps). In contrast, the full cortical analysis estimates modulations along both time and frequency axes in addition to an integrated view of the two axes including orientation information The analysis based on the separable model achieved a correlation coefficient of 0.83 (Table 3.2).

Second, we further reduced the separable spectro-temporal space by analyzing the modulation content along both time and frequency without maintaining the distribution along the tonotopic axis. This was achieved by simply integrating the modulation features along the spectral axis thus exploring the global characteristic of modulation regardless of tonotopy (Figure 3.7C, rightmost panel). This representation is somewhat akin to what would result from a 2-dimensional Fourier analysis of the auditory spectrogram. This experiment yielded a correlation coefficient of 0.70 (Table 3.2), supporting the value of an explicit tonotopic axis in capturing subtle difference between instruments.

Next, we addressed the concern that the mere number of features included in the full cortical model enough to explain the observed performance. We therefore undersampled the full cortical model by employing only 6 scale filters; 10 rate filters and 64 frequency filters by coarsely sampling the range of spectro-temporal modulations. This mapping resulted in a total number of dimensions of 3840; to be comparable to the 4224 dimensions obtained from the separable model. We then performed the dimensionality reduction to 420 dimensions, similar to that used for the separable analysis discussed above. The correlation obtained was 0.86; which is better than that of the separable spectro-temporal model (see Figure 3.8). This result supports our main claim that the coverage provided by the cortical space allows extracting specific details in the musical notes that highlight information about the physical properties of each instrument; hence enabling classification and recognition.

Finally, we examined the value of the kernel-learning compared to using a simple Euclidian L2 distance at various stages of the model (e.g. peripheral model, cortical stage, reduced cortical model using tensor singular value decomposition). Table 3.2 summarizes the results of this analysis along various stages of the model shown in Figure 3.2. The analysis revealed that the kernel-based mapping does provide noticeable improvement to the predictive power of the model but cannot -by itself- explain the results since the same technique applied directly on the spectrum only yielded a correlation of 0.74.

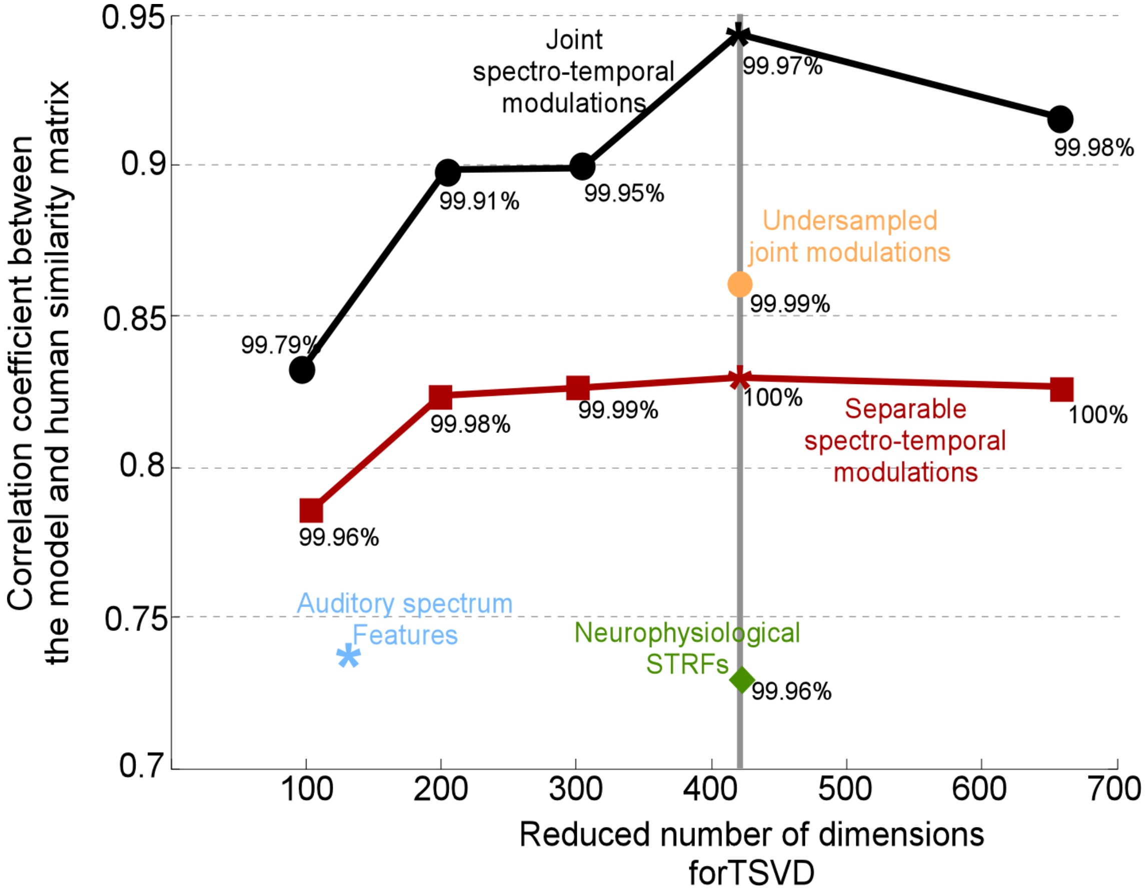

Figure 3.8: Correlation between human and model similarity matrices as a function of reduced feature dimensionality. In each simulation, a cortical representation is projected onto a lower dimensional space before passing it onto the SVM classifier. Each projection maintains a given percentage of the variability in the original data (shown adjacent to each correlation data point). We contrast performance using full cortical model (black curve) vs. separable spectro-temporal modulation model (red curve). The empirical optimal performance is achieved around 420 dimensions; which are the parameters reported in the main text in Table 3.2.

3.5 Application to Musical Performances

As we have seen previously, models learnt on a data base of musical notes are able to generalize well to unseen examples. It is desirable, however, to have a system that could also generalize to continuous musical performances by artists. Such musical recordings consist of sequences of notes and chords. This poses a challenge when a system trained on isolated notes is used to classify continuous music recordings due to the difference in temporal characteristics. We can either ignore the presence of note boundaries and uniformly window the musical recording to note-sized segments or identify regions of single note occurrences and segment them. The remainder of this section we describe how we implement both these approaches.

3.5.1 Note Extraction

Traditionally the task of note extraction has been morphed into the task of detecting note onsets and then labelling the region between two onsets as a note. Onset detection involves evaluating the given audio signal using an onset detection function and applying certain selection criteria to decide the onset times. Onsets are caused by a break in the steady state nature of a note. Thus previous onset detection attempts have tried to use phase deviation features which detect the departure from steady state behaviour [127] [128]. However, a note is also characterized by a region of relatively steady pitch and significant harmonicity level. We exploit this fact and propose a potential alternative to extract notes that would be able to identify regions of stable pitch frequency and high harmonicity. This method described in detail below performs better for instrument recognition when compared to some of the methods from [128] based on weighted phase deviation and rectified complex domain features on a small set of data. This is due to the fact that onset based approaches are sensitive to signal level characteristics which are easily affected by changing conditions like recording instruments and environments. However, the proposed method for note extraction performs reliably in varying conditions by using high level information like pitch and harmonicity estimates.

Continuous music recording consists of a series of notes and chords played in succession. To avoid the artifacts that occur from analyzing multiple notes, we segment continuous recordings into individual notes based on estimating pitch and its harmonicity. We estimate the pitch using a model based on a template matching proposed by Goldstein et al. [129]. We compare the spectrogram slice of the audio at any given time to templates generated at different pitches. These templates are models of the auditory spectrum of a generic note at a particular pitch value. The pitch of the best matching template is then assigned to be the pitch of the sound at that time. For this purpose, we chose the pitch template $T(f; f_p)$ to be a cosine function modulated by a Gaussian envelope(see Fig. 3.9 a) repeated at the harmonics of a

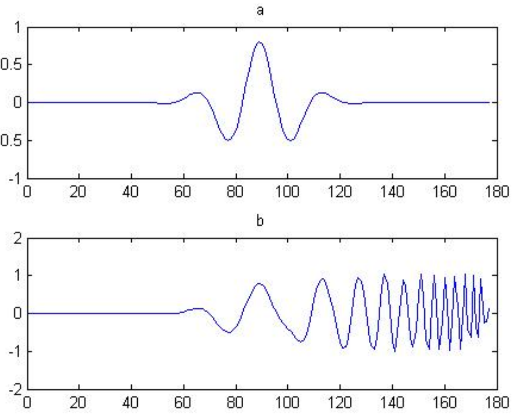

Figure 3.9: a. An example of the base function used to create the template b. An example of the template for a particular pitch frequency.

particular pitch frequency f_p as given by Eq 3.8.

$$rClT(f;f_p) \quad = \quad \sum_n 2e^{-\left(\frac{f-nf_p}{\alpha\theta(n)}\right)^2} cos(2\pi\frac{\theta(n)}{\beta}(f-nf_p)) \tag{3.8}$$

where $\theta(n) = 1 + 0.7 * n$ is a shrinkage factor, $\alpha = 18.28$ and $\beta = 26$ are constants. We then calculated the match in terms of correlation value $H(t;f_p)$ and used the maximum value to calculate $P(t)$, the pitch for that time instant, as shown in Eqs. 3.9, 3.10.

$$rClH(t;f_p) \quad = \quad corr(y(f,t),T(f;f_p)) \tag{3.9}$$

$$P(t) \quad = \quad arg\max_{f_p} H(t;f_p) \tag{3.10}$$

$$S(t) \quad = \quad \max_{f_p} H(t;f_p) \tag{3.11}$$

where $y(f,t)$ is the spectrogram and *corr* is the Pearson's correlation coefficient. The harmonicity $S(t)$ calculated as in Eq. 3.11 indicates the degree of match to the template at the selected pitch value. We know that boundary between 2 successive notes can be identified by the change in the pitch of sound. We also know that due to the overlap of notes at the note boundaries the harmonicity of the sound at the boundary is not high. We exploit these two facts and decide the note boundaries when both these criteria are met as described below.

We selected the potential note boundaries based on $P(t)$ as those time instances when the following condition is met:

$$Cabs(P(t) - \tilde{P(t)}) \geq \tau_1 \tag{3.12}$$

$$\tilde{P(t)} = mode(\{P(t-w), P(t-w+1)...P(t)\}) \tag{3.13}$$

where $\tau_1 = 0.2$ and $w = 30ms$.

We select the potential note boundaries based on $S(t)$ (which is normalised to be 0 mean and unit variance) when the following criteria are satisfied.

$$CS(t) \leq S(k) \ for \ all \ k : t - w \leq k \leq t + w \tag{3.14}$$

$$S(t) \leq \frac{\sum_{k=n-mw}^{n+w} S(k)}{mw + w + 1} - \tau_2 \tag{3.15}$$

$$S(t) \leq g_\mu(t-1) \tag{3.16}$$

where $m = w = 30ms$, $\tau_2 = 0.3$, and $g_\mu(t-1)$ is a thresholding function given by

$$Cg_\mu(t) = min(S(t), \mu g_\mu(t-1) + (1-\mu)S(t)) \tag{3.17}$$

Finally the note boundaries are selected as those times where the potential note boundaries based on both the pitch and harmonicity agree (with a tolerance of 40ms).

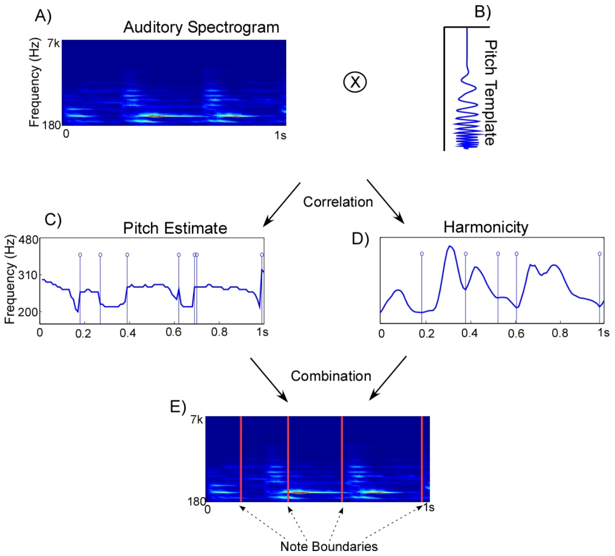

Figure 3.10: The note extraction scheme is depicted here. An example of spectrogram (A) of a piano audio segment containing 4 notes which is convolved with a pitch template (B) to yield the pitch estimate (C) and harmonicity (D) along with the candidate onset points. Finally the note boundaries are depicted in red (E) where the candidates coincide.

This method not only has the advantage of looking at high level features like pitch and harmonicity but also combines the information from these two methods to get reliable candidates for note boundaries as shown in the example in Fig. 3.10.

3.5.1.1 Choice of Notes

We created a new database of commercially available CDs of solo instrumental music performances recorded in studios for the six instruments mentioned above. This database contains, on average two hours of data per instrument. All the sounds were downsampled to 16kHz and pre-emphasized with a FIR filter with coefficients [1 ,-0.97]. The goal of this experiment was to test a model trained on isolated notes

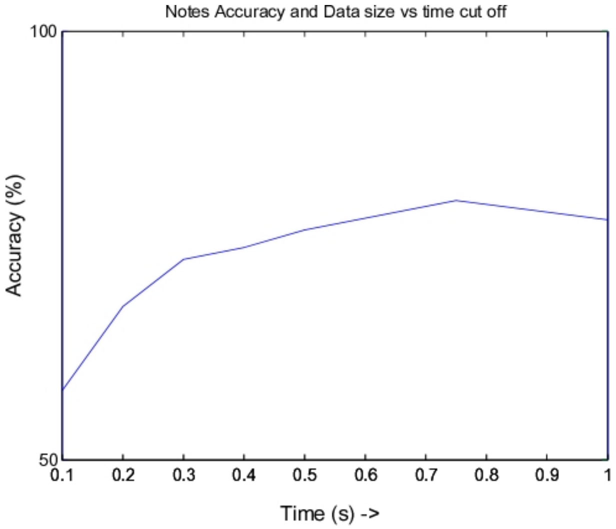

Figure 3.11: Note accuracy as a function of the time cutoff τ_n

from RWC database against notes extracted from such recordings. One significant source of error could be due to the fact that continuous recordings are not always a sequence of notes. For example, a piano performance can have chords which leads to the extraction of many small segments during note extraction due to the ambiguity of pitch estimate. To remove this source of error, we use only those notes extracted which are longer in time than a threshold τ_n. We tested for various values of τ_n experimentally by testing against a system trained on RWC database and the results are shown in Fig 3.11. Based on these results, we chose the optimal value of 750ms as the value of τ_n for all future experiments.

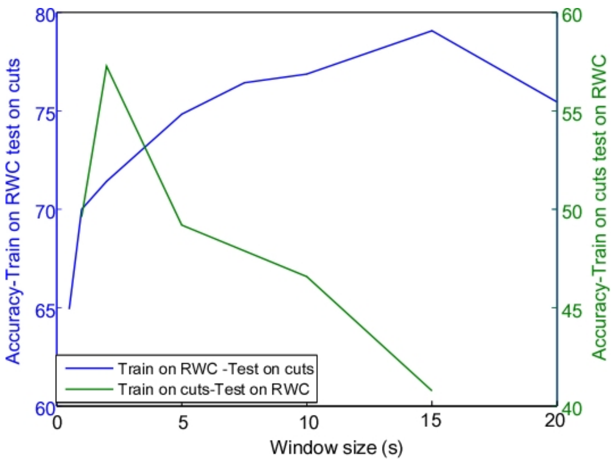

Figure 3.12: Accuracy for uniform windowing experiments i) and ii) as a function of the time cutoff τ_w

3.5.1.2 Short Term analysis

As a computationally less expensive alternative to note extraction, we used uniform windowing of solo music, where each recording was segmented into non-overlapping regions of duration τ_w. This method ignores the occurrences of notes or chords and treats each segment equally. The time duration, τ_w, for this window is a parameter which will control the degree of mismatch between the two databases. To estimate the optimal value of τ_w, we ran 2 sets of experiments: i) train on notes from RWC and test on uniform windowed segments from solo music database ii) train on uniform windowed segments from solo music database and test on RWC notes. There is a tradeoff between these results and we choose 2s as the value for τ_w for all future experiments as this seems to be quite important for experiment ii) as showing in Fig 3.12.

3.5.2 Cross-domain testing

We train 3 different classifiers on features extracted from the RWC notes, notes extracted from the solo music database and uniform window segments also from the solo music database. We then test these 3 classifiers again all 3 feature sets. This is done using k-fold cross validation technique with k=10. The results for these are shown in table 3.3.

Table 3.3: The cross testing results.

Train \Test	RWC	Notes	Windows
RWC	98.5% ± 0.2%	78% ± 2.1%	71% ± 1%
Notes	44.7% ± 0.9%	97.7% ± 0.6%	93.4% ± 0.5%
Windows	58.5% ± 1.5%	97.3% ± 0.5%	96.9% ± 0.4%

The high classification accuracy for the cross testing between the note extraction technique and uniform windowing technique on the solo music database tells us that there was a high degree of agreement between these methods. The higher accuracy for the note extraction technique as compared to the uniform windowing technique when tested against a classifier trained on RWC notes indicates that note extraction was better at reducing the difference between the datasets (as expected because the RWC dataset also has isolated notes). Finally the low classification accuracy for the classifiers trained on the feature sets derived from solo music database when tested on RWC database indicates that RWC database is a more generalized database with much more variance in the data as compared to the solo music database. Thus this model was limited and could not generalize to the wide range of variations present in RWC database.

3.5.2.1 Model Adaptation

A key issue when testing systems on out-of-domain data was the difference in statistical distributions of the different datasets as was demonstrated previously. These differences could have arisen due to many factors such as, differences in recording instruments, room acoustics, channel noise etc. Many adaptation techniques have been proposed in the literature to adapt existing machine learning models using a small set of data from the out of domain dataset. Here we assume that at least a limited amount of data is available to adapt a trained model. In order to adapt the SVM boundaries, we maximized the margin of the new data points using an objective function and learning mechanism similar to the original training of the model [130]. We used this technique to adapt a system trained on RWC database using 200 notes per class from the solo music recordings. The resulting improvement in performance from 80.27% to 86.6% indicated that the availability of limited data could be exploited to adapt the model to different conditions.

3.6 Discussion

This study demonstrates that perception of musical timbre could be effectively based on neural activations patterns that sounds evoke at the level of primary auditory cortex. Using neurophysiological recordings in mammalian auditory cortex as well as a simplified model of cortical processing, it is possible to accurately replicate human perceptual similarity judgments and classification performance among sounds from a large number of musical instruments. Of course, showing that the information is available at the level of primary auditory cortex does not imply that all neural correlates of sound identification will be found at this level. Nevertheless, it suggests that the spectro-temporal transforms as observed at this stage are critical for timbre perception. Moreover, our analysis highlights the ability of the cortical mapping to capture timbre properties of musical notes and instrument-specific characteristics regardless of pitch and playing style. Unlike static or reduced views of timbre that emphasize three or four parameters extracted from the acoustic waveform, the cortical analysis provides a dynamic view of the spectro-temporal modulations in the signal as they vary over time. A close examination of the contribution of different auditory features and processing stages to the timbre percepts highlights three key points.

First, neither the traditional spectrum nor its variants (e.g. average auditory spectrum [32]) are well-suited to account for timbre perception in full. According to our simulations, these representations encode the relevant spectral and temporal acoustic features too implicitly to lend themselves for exploitation by classifiers and other machine learning techniques. In some sense, this conclusion is expected given the multidimensional nature of the timbre percept compared to the dense two-dimensional spectrogram; and is in agreement with other findings from the literature [93].

Second, when considering more elaborate spectro-temporal cortical representations, it appears that the full representation accounts best for human performance. The match worsens if instead marginals are used by collapsing the cortical representation onto one or more dimensions to extract the purely spectral or temporal axes or scale-rate map (Figure 3.8, Tables 3.1 and 3.2). This is the case even if all dimensions are used separately, suggesting that there are joint spectro-temporal features that are key to a full accounting of timbre. While the role of both purely spectral and temporal cues in musical timbre is quite established [86], our analysis emphasizes the crucial contribution of a joint spectro-temporal representation. For example, FM modulations typical of vibrato in string instruments are joint features that cannot be easily captured by the marginal spectral or temporal representations. Interestingly, acoustical analyses and fMRI data in monkeys suggest that the spectro-temporal processing scheme used here may be able to differentiate between broad sound categories (such as monkey calls vs. bird calls vs. human voice), with corresponding neural correlates when listening to those sounds [131].

Third, a nonlinear decision boundary in the SVM classifier is essential to attain the highest possible match between the cortical representation and human perception. Linear metrics such as L2 are less optimal, indicating that the linear cortical representation may not be sufficiently versatile to capture the nuances of various timbres. The inadequacy of the linear cortical mapping has previously been described when analyzing neural responses to complex sounds such as speech at the level of auditory cortex [110,111,113]. In these cases, it is necessary to postulate the existence of nonlinearities such as divisive normalization or synaptic depression that follows a linear spectro-temporal analysis so as to account fully for the observed responses. In the current study, the exact nature of the nonlinearity remains unclear as it is implicitly subsumed in the Gaussian kernels and subsequent decisions.

We then attempted to extened this model to continuous musical recordings. We showed that a constant windowing approach was not able to capture all the nuances present at the note boundaries. We proposed a note extraction scheme based on a template matching approach which is believed to be biologically plausible. This approach along with model adaptation was shown to be the most beneficial in cross testing of models trained on musical notes and tested on musical performances.

In summary, this study leads to the general conclusion that timbre percepts can be effectively explained by the joint spectro-temporal analysis performed at the level of mammalian auditory cortex. However, unlike the small number of spectral or temporal dimensions that have been traditionally considered in the timbre literature, we cannot highlight a simple set of neural dimensions subserving timbre perception. Instead, the model suggests that subtle perceptual distinctions exhibited by human listeners are based on 'opportunistic' acoustic dimensions [100] that are selected and enhanced, when required, on the rich baseline provided by the cortical spectro-temporal representation.

Chapter 4

Multi-dimensional Representations for Soundscapes and Role of Attention

4.1 Introduction

An auditory object is often equated to the sound produced by a single source [132]. While the correspondence between the two is not always a one-to-one mapping, the soundscape incident on a listener generally consists of multiple auditory objects that constitute the acoustic scene. Identifying an auditory object is not a trivial task, especially since each object can present itself in a multitude of variations. For example, the blast of a car horn can differ depending on the make, the speed of the vehicle, and also the distance of the vehicle from the listener. Subsequently, this makes the identification of a collection of auditory objects (i.e. the acoustic scene) significantly harder. To add to this complexity, the nature and number of acoustic objects in a typical scene change over time. In a street, the sounds coming from passing cars can blend every now and then with speech from pedestrians or music from street artists. Changing scenarios add a new dimension of difficulty to the task of acoustic scene classification.

Attempts at automatic acoustic event and scene classification have typically followed the path of extracting short term features from waveforms and learning the statistics of these features to later classify an unknown example. Mel Frequency Cepstral Coefficients (MFCC), filterbank energies or Perceptual Linear Prediction coefficients (PLP) have been popularly used as features for this task [23–25]. They are often complimented with other low level features like zero crossing rate, short time energy, spectral flux, pitch, brightness and bandwidth [24, 26, 27] or are transformed to account for long term statistics [28]. Though short term spectral attributes coupled with low-level features have been quite successful in a number of applications, studies have also shown that they are limited in capturing the full range of information relevant for acoustic scene recognition; and that joint local modulations in energy along both time and frequency are able to better capture the qualities of acoustic scenes [29]. This rich modulation space builds on neurophysiological studies in the mammalian auditory system indicating that neurons at the level of auditory cortex respond to local joint spectral and temporal modulation in the signal [30]. This biological analysis can be viewed as mapping sound onto a high dimensional feature space which captures the detailed variations of the spectral profile and its temporal variations, as a basis for representing acoustic events.

This sensory mapping is complemented with cognitive mechanisms, most notably task-driven attention, which allows us to isolate and recognize objects of interest amidst other competing sound events [36]. Neurophysiological and brain imaging studies have shown that task-driven attention modulates the gain of sensory cortex responses to highlight features of interest [133, 134]. Attention has been argued to act as a Bayesian prior representing distribution of beliefs acting as gating mechanism to reduce uncertainty, to increase signal-to-noise ratio or to refine perceptual inference around some goal-specific point in sensory space [135]. In addition, attention is also believed to modulate cognitive and decision-making frontal areas of the brain, most notably prefrontal cortex [136]. Psychoacoustic evidence also supports the premise that attention operates at multiple levels, be it feature-based or object-based levels of representation [40–42, 137]. Motivated by these observations, the current study attempts to develop a scheme that incorporates attentional mechanisms in a model for scene recognition for multi-source environments. The model focuses on attentional processes operating at the level of both sensory representation and cognitive decisions.

4.2 Methods

4.2.1 Experimental Setup

4.2.1.1 Data

All the experiments are conducted on the BBC Sound Effects Database [138] which contains 18 scene classes. We selected 12 scenes from these as listed in Table 1 based on the number of recordings which in total resulted in 47 hours of data. The recordings which are at 44.1kHz sampling rate are down sampled to 16kHz and pre-emphasized using a filter with coefficients [1 -0.97]. These recordings are then randomly divided into training and testing sets in a 9:1 ratio. To simulate a multi-source environment (ME) we further mix the recordings in the test set with recordings from other classes at different target to masker ratios (TMR) from -20dB to 20dB in steps of 5dB. In addition to this, as an Adaptation Set(AS) we create 180 randomly chosen 1s segments for each class from the training set and simulated the multisource environment conditions similar to the testing set as explained above.

4.2.1.2 Testing

All the models will be trained on the original training dataset and some experimental conditions are allowed to use the Adaptation set to adapt different parameters. The model is then tested on the recordings in the ME set and is asked to detect the presence of the target class. We consider a measure that not only considers the hit rate (HR) but also the false alarms (FA) to ensure that the model is not biased towards recognizing the target class. This measure is the Dprime measure evaluated as $d' = Z(HR) - Z(FA)$ where $Z(.)$ is the inverse cumulative distribution function of a standard normal distribution. Consider an example confusion matrix show in Table 4.1 which was a result of evaluating the system with class 2 designated as the target class. Then the Hit Rate is calculated as $HR2 = X22/(X21+X22+X23)$. The False Alarm rate would be calculated as $FA2 = (X12 + X32)/(X11 + X13 + X31 + X33)$. Thus we need an improvement in not only correct recognition of the target when it is present but also to recognize the absence of target to get an improved Dprime score.

Table 4.1: An example confusion matrix with three classes.

True Label \Predicted Label	1	2	3
1	$X11$	$X12$	$X13$
2	$X21$	$X22$	$X23$
3	$X31$	$X32$	$X33$

4.2.1.3 Baseline

We compare the proposed sensory based feature representations with two state of the art representations for auditory scene analysis : 1) MFCC [23] 2) Opensmile features [139] . For the MFCC features we use a Hamming window length of 25ms with an overlap of 15ms and 13 dimensional MFCCs are calculated and finally the C0 component is ignored. The mean, standard deviation and skew is calculated for these 12 coefficients over the duration of the segment and concatenated resulting in 36 dimensional feature vectors.

For the Opensmile features we use the opensource toolkit [140] with the configuration set to emo_large.conf feature set. This scheme extracts four main basic features sets namely spectral, cepstral, energy and voicing features. The derivative and acceleration coefficients of these basic features are concatenated. These short term features are then combined into a single feature for the entire duration of the considered segment using 39 statistical functionals like mean, standard deviation, quartiles, percentiles etc resulting in a 6669 dimensional feature vector. These features are then reduced to a 120 dimensional feature vector using SVD technique to keep 99.99% of the variance.

4.2.2 Scene Analysis Model

The proposed model for scene analysis (Figure 4.1) can be divided into the sensory based feature representation stage and the object representation stage. We furthermore implemented adaptation strategies to readjust both these stages of the model to a given target class.

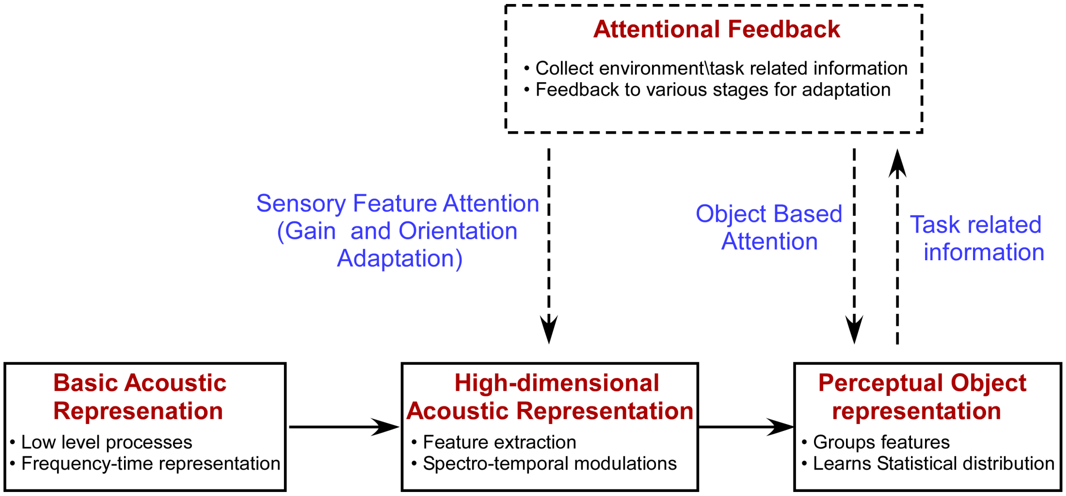
Attentional Feedback
• Collect environment\task related information
• Feedback to various stages for adaptation
Sensory Feature Attention (Gain and Orientation Adaptation)
Object Based Attention
Task related information
Basic Acoustic Representation
• Low level processes
• Frequency-time representation
High-dimensional Acoustic Representation
• Feature extraction
• Spectro-temporal modulations
Perceptual Object representation
• Groups features
• Learns Statistical distribution

Figure 4.1: Schematic of the proposed scene analysis model. The basic and high-dimensional acoustic representation modules project the acoustic stimuli into a robust, discriminative high dimensional representation as described in 4.2.2.1. The perceptual object recognition module collects the statistics of sound objects and learns to discriminate classes as described in 4.2.2.2. The attentional feedback module collects information related to the target and the environmental information related to the task at hand. It then relays back information to influence the gain and orientation of the high dimensional feature extraction stage and also the statistical models in the object representation stage. These influences tune the system to better detect the presence of the target source sounds.

Table 4.2: List of Classes used and number of recordings and amount of data in minutes for each class.

Class	Number of Recordings	Minutes
Transportation	559	781
Animals	426	597
Foley	220	133
Industry	140	181
Humans	129	326
Emergency	119	142
Sports	77	156
Technology	77	60
Water	60	178
SciFi	58	21
Households	48	83
Weather	41	155

4.2.2.1 Sensory Feature Representation

We implemented a feature representation scheme broadly based on the processing of the sound in the mammalian auditory system. This scheme is further divided into 3 stages. The first stage models the transformations from the periphery to the cochlear nucleus and is given by Equation 1.1 in Section 1.2. The smoothing filter $\mu(t;\tau)$ is applied with a time constant of $\tau = 4ms$. The resulting time frequency representation $z(f,t)$ from the second stage is referred to as the auditory spectrogram. The third stage of processing mimics the behavior of the primary auditory cortex. The neurons at this stage are sensitive to various features like frequency, pitch, temporal variations, spectral modulations, joint spectrotemporal variations etc. Thus these neurons are effectively acting as feature detectors for the higher stages of processing in the auditory cortex. We modeled these as a bank of 2d Gabor filters (GF(.)) which are sensitive to joint temporal and spectral variations (denoted by rates r and scales s respectively) as given in Equation 1.3 in Section 1.2. These filters are linear approximations to the receptive field shapes found in the primary auditory cortex [34, 109]. The filtering operation can be viewed as a two dimensional convolution with the auditory spectrogram followed by time averaging of the absolute response as shown in Equation 4.1, resulting in a 3 dimensional tensor representation of the cortical features $CF(f,s,r)$. The averaging is done over non overlapping 1s windows for each recording. We consider the Gabor filters at 10 upward and 10 downward rates in equal logarithmic steps from 2Hz to 215Hz; and at 11 scales from 0.25 cycles per octave to 8 cycles per octave at uniform steps in the logarithmic scale. The features are reduced to a size of 120 features (x) retaining 99% of the variance using the Tensor SVD dimensionality reduction method [101].

$$CF(s,r;f) = \sum_t |z(f,t) \otimes_{t,f} GF(s,r;f,t)| \tag{4.1}$$

4.2.2.2 Object Representation

The features from the training data for each class are used to learn a Gaussian mixture model (GMM) representation which models the distribution of the features. We use 128 mixtures for each class with diagonal covariance matrices. During testing given a recording we extract $x(t)$, the time series of reduced features as explained above and the class with the highest posterior probability under the learnt GMM is chosen as the label($\tilde{l}$) for the recording as shown in Equation 4.2.

$$\tilde{l} = arg \max_{l} P(z(t)|l) \tag{4.2}$$

4.2.3 Models of Attention at the Sensory Feature Representation stage

As discussed in the previous section, task driven attentional effects at the sensory representation stage can be seen as a boosting or gain of the relevant feature detectors or as a shape change to detect new relevant features. We describe below how we incorporate these in our model in order to improve the performance in detecting the given target.

4.2.3.1 Gain Adaptation

Given a target to attend to, we adapt the gains of the Gabor filters using prior knowledge of the target. Here we use the mean activity pattern (M_l) from the entire training data as the prior knowledge to adapt the gains. M_l is normalized to be between 0 and 1. We limit the level of adaptation using a gain control parameter α. The gains (g_l) are limited to be in the range of $1-\alpha$ to 1 as shown in Equation 4.3. We then simply weight the output of the filtering operation with these gain coefficients as in Equation 4.4. This operation should suppress those features which were not active during training and retain only the active features.

$$g_l(f,s,r) = (1-\alpha) + M_l(f,s,r) * \alpha \tag{4.3}$$

$$\hat{C F}_l(s,r;f) = g_l(f,s,r) * CF(s,r;f) \tag{4.4}$$

4.2.3.2 Orientation Adaptation

In order to attend to the target class (l), we propose a mechanism to fine tune the filter shapes in such a way that the features of the target class are enhanced. Here we allow the filter to change its shape only by changing the orientation of the main lobe (Figure 4.2) within a fixed limit of 3 degrees. The orientation is calculated as $\theta = \arctan(s/r)$. Using the adaptation set for the target class, we calculate the best orientation for each filter, which results in most similar maximum activity to the mean activity of the target class, at the output of the filters on average, as shown in Equation 4.5. We then use these pairs of filter parameters to design the Gabor filters and extract features as in Equation 4.1. Here $theta_0$ represents the original orientation, $th\hat{e}ta_l$ represents the optimized orientation.

$$\hat{\theta}_l = arg \max_{|\theta-\theta_0|\leq 3} \sum_f (CF_l(s,r;f) - M_l(s_0,r_0;f))^2 \tag{4.5}$$

Thus by allowing the filters to fine tune their orientation to the target in the multisource setting the detection performance should be enhanced.

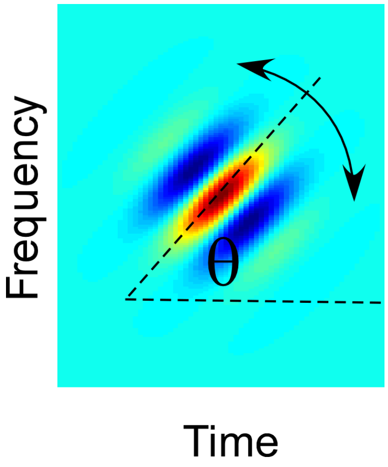

Figure 4.2: An example of a Gabor Filter. The orientation θ for the filter is calculated from the slope of the main lobe of the Gabor filter. Orientation Adaptation allows this orientation to be changed within a limit of 3 degrees such that the similarity to the target mean representation is maximized.

4.2.4 Models of Attention at the Object Representation stage

Attending to a particular target is known to induce adaptation at higher stages of processing in the auditory system, which deal with object representations [41,136]. In order to model this phenomenon we use a strategy of adapting the guassian mixtures using MAP adaptation technique. MAP adaptation has been proven to be useful for many applications like Image segmentation [141], EEG verifications [142], Speaker Verification [143] etc. We use the adaptation set for the target class to calculate the reduced features X. Using these features, we adapt the GMM parameters (φ), which contain the means and the prior probabilities of the target class mixtures, as shown in Equation 4.6. Here $\gamma = (1+\beta)^{-1}$ where β is the relevance parameter which controls the degree of adaptation. Increasing values of β leads to more reliance on the new data. This mechanism of adaptation should adjust the target class model to better represent the distribution of the target in the multisource testing condition.

$$\hat{\varphi} = arg\max_{\varphi} P(X|\varphi)^{1-\gamma}P(\varphi)^{\gamma} \tag{4.6}$$

4.3 Results

In order to test the performance of the proposed system, we train the GMM model on the clean training set, and further adapt the system to the considered target class using the Adaptation set as explained in Sections 4.2.3, 4.2.4. This system is then tested in the ME condition using the artificially generated ME set to generate a confusion matrix for each SNR condition. We then compute the d' metric for each confusion matrix generated. We repeat this process for all the 12 classes as the target class. The results for the proposed system, Opensmile baseline and the MFCC baseline system are given in Table 4.3. We observe an doubling and tripling in the dprime result of the proposed system when compared to the OpenSmile baseline and MFCC baseline respectively. This remarkable improvement shows the ability of the proposed system to better identify the target sources in the multisource environment. In order to better assess the contribution of the different adaptation techniques of the proposed system, we test the different components of the system below.

Table 4.3: Dprime evaluation results for the proposed system with and without adaptation and the baseline systems.

System	Dprime
Proposed System	1.15
Proposed System without Adaptation	0.52
Opensmile features	0.51
MFCC features	0.36

4.3.1 Gain Adaptation

The proposed gain adaptation technique uses prior knowledge of the target class mean activation pattern as described before in Section 4.2.3.1. This method boosts the output of those GF filters whose mean activitty is high in the target class. Thus in the ME testing condition we boost the target representation and supress the interference sources. In order to test this system we evaluate a system without Orientation Adaptation. Within such a system we further test for - with and without Object Adaptation conditions. The results for these testing conditions are shown in Table

4.4. The 48% relative improvement of the Gain Adaptation system with respect to the Proposed System without any Adaptation techniques shows that the proposed approach to suppressing the background distractions while enhancing the target representation is effective. Moreover applying the Object Adaptation technique enhances the Gain Adaptation by a further 45% relative improvement. This complimentary improvements by the adapation techniques at the Feature representation and Object representation stages further enforces the ability of the Gain Adaptation technique to boost the target representation.

Table 4.4: Dprime evaluation results for Gain Adaptation system with and without Object Adaptation.

System	Dprime
Proposed System	1.15
Gain Adaptation & Object Adaptation	1.12
Gain Adaptation	0.77
Proposed System without Adaptation	0.52

4.3.2 Orientation Adaptation

The propsed Orientation Adaptation technique uses prior knowledge about the target class and adapts the orientation of the Gabor Filters using the AS examples to fine tune the orientation of the main lobe to maximize the similarity of the response to the target class. Thus we enhance the representation of the target class and supress the representation of the other sources in the ME condition. To test this hypothesis we evaluate the proposed system without Gain Adaptation, both with and without Object Adaptation. The results of these evaluations can be found in Table 4.5. The 19% relative improvement is obtained by the Orientation Adaptation technique when compared to the proposed system without any Adaptation applied. This shows the ability of the Orientation Adaptation technique to enhance the target source representation in the presence of distracting background sources. A further 27% relative improvement is achieved by coupling the Object Apatation and Orientation Adaptation techniques. This shows that by further adapting the statistical representation of the Objecte Representation stage, we are able to further enhance the separation of the target class induced by Orientation Adaptation.

Table 4.5: Dprime evaluation results for Orientation Adaptation system with and without Object Adaptation.

System	Dprime
Proposed System	1.15
Orientation Adaptation & Object Adaptation	0.79
Orientation Adaptation	0.62
Proposed System without Adaptation	0.52

4.3.3 Object Adaptation

The propsed Object Adaptation technique uses the AS examples to change the statistics of the Object Representation stage to better fit the multisource evaluation scenario. We use MAP adaptation technique to move the means and the prior probabilites of the Gaussian Mixtures in the Object Representation stage to the new statistics in the ME setting. In order to test the Object Adaptation stage we evaluate the proposed system with and without Object Adaptation. We also evaluate the proposed system with and without Gain and Orientation Adaptation. The results of these tests can be found in Table 4.6. We see a 36% relative improvement by applying the Object Adaptation technique as compared to a system without any Adaptation. However, when Object Adaptation technique is applied to a system with Gain and Orientation Adaptation we see a 47% relative improvement in performance. Thus we can see that 1) the Object Adapation technique is effective at modifying the Object Representation to match the changed statistics in the ME testing condition 2) the Object Adaptation technique performs better in conjunction with Gain and Orientation Adaption showing that the enhancement of target representation is essential to best identify the target in multisource setting. All the results above are assimilated in the Figure 4.3 to show the progression of improvements achieved by the various adaptation techniques.

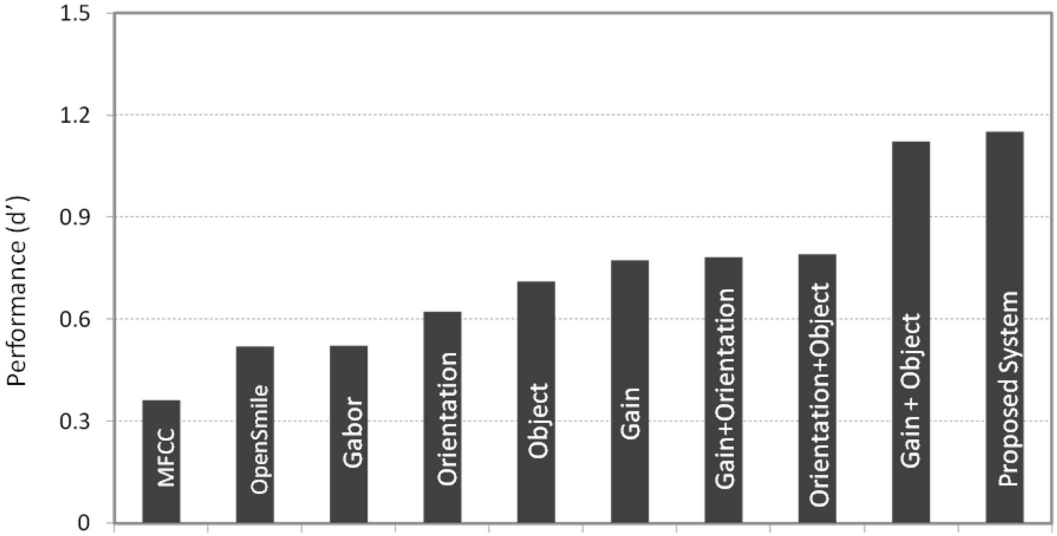

Figure 4.3: Overall results on the ME testing condition. The proposed system, baseline systems and the various adaptations utilized in the proposed system.

Table 4.6: Dprime evaluation results for Proposed system with and without Object Adaptation. We also evaluate the Object Adaptation without Gain and Orientation Adaptation.

System	Dprime
Proposed System	1.15
Gain & Orientation Adaptation	0.78
Object Adaptation	0.71
Proposed System without Adaptation	0.52

4.4 Discussion

Automatic systems for auditory scene recognition or sound object recognition consist of two main components - feature extraction and statistical modelling. Existing studies focus on improving the feature extraction and building improved statistical models [23, 144]. However all of these systems are feed forward in nature [24, 26]. Multiple studies attempt to adapt the system to changing evironmental conditions. This is achieved by the use of statistical techniques which adapt the statistical models [141–143]. Neurophysiological studies however show that top-down attentional mechanisms affect the sensory cortex responses and the higher decision making stages along the auditory pathway [133, 134, 136]. It is important to however note that such mechanisms differ from the more well known bottom-up attentional mechanisms. Such bottom-up mechanisms are known to be feed-forward and involuntary and come into action when a salient event captures your attention. For example when a lound bang captures your attention in an otherwise normal scenario. Top-down attentional mechanisms are voluntary and come into effect when the subject predecides to pay attention to a target. For example when we are trying to pay attention to speech in the middle of a noisy street.

In this study we use a representation of sound which was not only proven to provide high degree of seperability between different classes of audio but also provides a convinient mechanism to incorporate top-down task driven attention [145]. Here we proposed new techniques to improve the recognition of target classes in the presence of multiple sources. The proposed system consists of two techniques to adapt the Sensory Feature Representation stage and a technique to adapt the Object Representation stage. At the Feature Representation stage we proposed a technique to change the gain of the Gabor Filters to enhance the target class represenation. We also proposed a techniqe to change the orienation of the main lobe of the Gabor Filter to fine tune the filters to the target class in the ME condition. At the Object Represenation stage we adapt the statistical model to better fit the changing statistics in the ME condition. We showed the effectiveness of the proposed techniques individually and in conjugation with each other at detecting the target class. Such an adaptation technique would be useful in various applications for mobile devices which are often used in various acoustic conditions with multiple sources.

In the future we aim to expand this work to build a system which can pay attention to a target class but also continually adapt to changing environmental conditions. We also intend to examine the ability of the system to choose the attentional mechanism and the degree of modulation to be used automatically depending on the target class, the current manifestation of the target source and the distracting background sources. Another direction of investigation would be to incorporate the ability of non-parametric shape changes of the Gabor filters to capture the target sounds.

Chapter 5

Conclusion

5.1 Thesis Overview

The human auditory perception system is exposed to a wide variety of acoustic stimuli, be it a well defined musical note, a friend speaking, or a complex acoustic scene like a bar or a busy fair. Humans are able to effortlessly analyse all these complex sounds and extract meaningful information from them. This suggests that a similar processing mechanism is sufficient to deal with a wide variety of acoustic sound classes. In this thesis we tested this hypothesis by proposing a high dimensional sound object representation mechanism which was motivated by the neurophysiological studies on the transformations in the auditory pathway. This representation captures the various modulations present in sounds by analyzing the sounds locally at multiple resolutions leading to a multi-dimensional representation. Such a high dimensional mapping projects different sounds to distinct regions leading to better discriminability. We then showed that this model is able to capture a wide variety of sound classes (speech, music, soundscapes) by applying it to the tasks of speech recognition, speaker verification, musical instrument recognition and acoustic soundscape recognition.

We first tested the proposed hierarchical sound object recognition system's ability to capture speech sound objects. The spectral characteristics of speech sound objects demonstrate a lot of intricate information which can be used to identify them. Thus we modified the proposed high dimensional representation to focus on the analysis of spectral characteristics in a detailed yet computationally-efficient manner. This approach captures the promient features of the spectrum varying from the broad variations (related to vocal tract shape and length) to finer modulations (related to harmonics and voice quality). We showed how this representation yields itself to capturing speech information content and the messenger/speaker information and how it can be fine tuned to these tasks. While the well known cepstral coefficients (MFCC) capture global modulations along the spectral slice, the proposed approach captures local modulations as various resolutions. We showed that while the two approaches are comparable in performance in clean conditions, the proposed approach outperforms a state-of-art system in mismatch noisy testing condition for both the speech recognition and speaker verification tasks.

The proposed sound object recognition system was then applied to the task of musical instrument recognition. We showed that using a model of cortical processing, we were able to accurately replicate the human perceptual similarity judgments and also were able to get a good classification performance on a large set of musical instruments. Our analysis showed that the proposed representation captures instrument identity regardless of pitch and playing style. From the perception matching experiments we learnt that 1) neither the traditional spectrum nor its variants were well-suited to account for timbre perception 2) the full cortical representaion is essential and that none of the marginals are sufficient 3) a non-linear decision boundary is essential in the SVM to get the best match and the presence of non-linearities at the cortical stage is well known. Moreover, we were able to extend this model to continuous musical recordings where we showed that a biologically plausible template based note extraction scheme along with model adaptation were successful in extrapolating the models learnt on notes to the new setting. In summary, this study leads to the general conclusion that timbre percepts can be effectively explained by the joint spectro-temporal analysis performed at the level of mammalian auditory cortex.

The previous approaches at classifying acoustic scenes reply on representations of the signal characteristics like zero crossing rate, loudness etc. in combination with spectral features like cepstral coefficients, perceptual linear prediction coefficients etc. In this thesis we showed that a simple set of features based on capturing the modulations from the time-frequency auditory spectrogram in a joint manner is sufficient to capture the identity of the acoustic scenes and outperform the previously reported techniques. However in complex soundscapes there are multiple sound sources producing sounds at the same time. Thus we saw the need to have a mechanism of selecting a target source from this mix of sources.

5.1.1 Model of Top-down Attention

The auditory system is well known to be a bi-directional system where, in the bottom-up direction the transformations on the stimuli from the environment occurs as the information travels upwards towards decision making areas, and in the top-down direction feedback information is applied on these lower level transformations in such a way to either enchance the representation of the target sounds or to better adapt to the changing environmental conditions. In such a system the top-down influences act in a complimentary manner enhancing the output of the bottom-up system. The previously proposed models in the literature for soundscape recognition follow a feed-forward topology which is static and does not adapt to different tasks. In this thesis we proposed a flexible framework where we were able to implement a top-down attention module which is complimentary to the proposed high dimensional acoustic feature extraction mechanism. This attention module is a distributed system operating at multiple stages. It affects the Sensory Feature Representation stage by adapting the gain and orientation of the filters. It also affects the Object Representation stage by moving the statistical models to better fit the changing environmental conditions. In summary, the attentional feedback module acts as a retuning mechanism, that adapts the same system to different tasks. We showed that such an adaptation mechanism is able to tremendously improve the performance of the system at detecting the target source in the presence of various distracting background sources.

5.2 Alternative Hypothesis

In this thesis we tested the hypothesis that similar representations are capable of representing various sounds belonging to diverse categories. However there could be alternate strategies employed by the auditory system in order to tackle different tasks. For example some studies show that there could be dedicated processing pathways for analyzing different classes of sounds [13]. This could mean that the representations along these different paths could indeed be distinct from each other. There is further evidence which shows that within speech there are distinct streams of processing for slow and fast varying components of speech [1]. The information from these streams could then be combined at a later stage of processing (See Figure 5.1). We have done some work incorporating this knowledge to provide improved enhancement of speech recognition by dividing the processing of modulations into different streams [146].

Alternatively, there could be different distinguishing features of sounds that are not captured by the proposed modulation analysis. For example, there is some evidence of template based matching in the auditory pathway that could be useful for

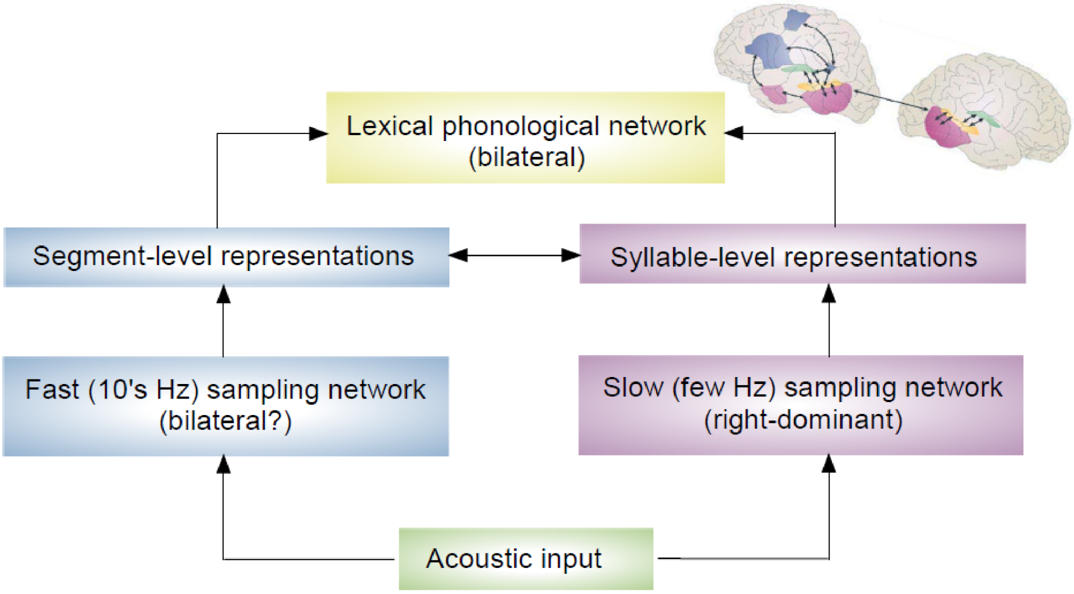
Lexical phonological network
(bilateral)
Segment-level representations
Syllable-level representations
Fast (10's Hz) sampling network
(bilateral?)
Slow (few Hz) sampling network
(right-dominant)
Acoustic input

Figure 5.1: Multiple streams of Processing. A distint pathways for processing the slow and fast varying components of speech. Adapted from [1]

pitch perception [147]. Note that we have also incorporated a template matching based approach for note extraction as described in Section 3.5.1. In the future we would like to investigate further into template based approaches for sound object recognition and possibly apply it in a different stream of processing which would act in a complimentary manner to the proposed high dimensional modulation analysis.

Bibliography

[1] G. Hickock and D. Poeppel, “The cortical organization of speech processing,” *Nature neurosc.reviews*, vol. 8, pp. 393–402, 2007.

[2] J. M. G. Gordon and J. W., “Perceptual effects of spectral modifications on musical timbres,” *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493–1500, 1978.

[3] S. McAdams, J. W. Beauchamp, and S. Meneguzzi, “Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters,” *The Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 882–897, 1999.

[4] R. D. Patterson, “The sound of a sinusoid: Time-interval models,” *The Journal of the Acoustical Society of America*, vol. 96, pp. 1419–1428, 1994.

[5] C. krumhansl, *Why is musical timbre so hard to understand?*, ser. structure and perception of electroacoustic sound and music. Amestrdam: Excerpta medica, 1989, pp. 43–53.

[6] S. McAdams, S. Winsberg, S. Donnadieu, G. D. Soete, and J. Krimphoff, “Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes,” *Psychological Research*, vol. 58, no. 3, pp. 177–192, 1995.

[7] J. M. Grey, “Multidimensional perceptual scaling of musical timbres,” *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.

[8] J. Burgoyne and S. Mcadams, “A meta-analysis of timbre perception using nonlinear extensions to clascal,” in *Computer Music Modeling and Retrieval. Sense of Sounds*, R. Kronland-Martinet, S. Ystad, and K. Jensen, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. A Meta-analysis of Timbre Perception Using Nonlinear Extensions to CLASCAL, pp. 181–202.

[9] S. Donnadieu, *Mental Representation of the Timbre of Complex Sounds*, ser. Analysis, Synthesis, and Perception of Musical Sounds. New York: Springer, 2007, pp. 272–319.

[10] A. Caclin, E. Brattico, M. Tervaniemi, R. Naatanen, D. Morlet, M. H. Giard, and S. McAdams, “Separate neural processing of timbre dimensions in auditory sensory memory,” *Journal of Cognitive Neuroscience*, vol. 18, no. 12, pp. 1959–1972, 2006.

[11] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, "Voice-selective areas in human auditory cortex," *Nature*, vol. 403, no. 6767, pp. 309–312, 2000.

[12] S. Uppenkamp, I. S. Johnsrude, D. Norris, W. Marslen-Wilson, and R. D. Patterson, "Locating the initial stages of speech-sound processing in human temporal cortex," *Neuroimage*, vol. 31, no. 3, pp. 1284–1296, 2006.

[13] J. W. Lewis, J. A. Brefczynski, R. E. Phinney, J. J. Janik, and E. A. DeYoe, "Distinct cortical pathways for processing tool versus animal sounds," *The Journal Of Neuroscience*, vol. 25, no. 21, pp. 5148–5158, 2005.

[14] A. M. Leaver and J. P. Rauschecker, "Cortical representation of natural complex sounds: effects of acoustic features and auditory object category," *Journal of Neuroscience*, vol. 30, no. 22, pp. 7604–7612, 2010.

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

[16] H. Hermansky, "Should recognizers have ears?" *Speech Commun.*, vol. 25, pp. 3–27, 1998.

[17] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans.Acoustic, Speech and Signal Process.*, vol. 27, pp. 113–120, 1979.

[18] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans.Acoustic, Speech and Signal Process.*, vol. 29, pp. 254–272, 1981.

[19] G. D. Cook, D. J. Kershaw, J. D. M. Christie, C. W. Seymour, and S. R. Waterhouse, “Transcription of broadcast television and radio news: the 1996 abbot system,” in *Proc. Int. Acoustics Speech and Signal Processing*, 1997, pp. 723–726.

[20] C. Chen and J. Bilmes, “Mva processing of speech features,” *IEEE Trans.on Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 257–270, 2007.

[21] ETSI, “Etsi es 202 050 v1.1.1 stq; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms,” 2002.

[22] P. Loizou, *Speech Enhancement: Theory and Practice.* (Boca Raton, FL): CRC Press, 2007.

[23] O. Kalinli, S. Sundaram, and S. Narayanan, “Saliency-driven unstructured acoustic scene classification using latent perceptual indexing,” in *IEEE International Workshop on Multimedia Signal Processing (MMSP 2009).* IEEE, 2009, pp. 478–483.

[24] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, “Non-speech audio event detection,” in *Proceedings of ICAASP’09*, april 2009, pp. 1973 –1976.

[25] X. Zhuang, X. Zhou, T. Huang, and M. Hasegawa-Jhonson, "Feature analysis and selection for acoustic event detection," in *Proceedings of ICAASP08*, 2008.

[26] A. Temko and C. Nadeu, "Classification of acoustic events using svm-based clustering schemes," *Pattern Recognition*, vol. 39, pp. 682–694, 2006.

[27] R. Cai, L. Lu, A. Hanjalic, H. Zhang, and L. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 14, no. 3, pp. 1026–1039, May 2006.

[28] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, ser. MIR '08. New York, NY, USA: ACM, 2008, pp. 105–112.

[29] K. Patil and M. Elhilali, "Goal-oriented auditory scene recognition," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[30] L. M. Miller, M. A. Escabi, H. L. Read, and C. E. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," *Journal of neurophysiology*, vol. 87, no. 1, pp. 516–527, 2002.

[31] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 3rd ed. Emerald Group Publishing Limited, 2008.

[32] X. Yang, K. Wang, and S. A. Shamma, “Auditory representations of acoustic signals,” *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[33] 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.

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

[35] K. M. and G. D., “Improving word accuracy with gabor feature extraction,” in *ICSLP-2002*, 2002, pp. 25–28.

[36] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, “Auditory attention–focusing the searchlight on sound,” *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[37] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, “Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex,” *Nature neuroscience*, vol. 6, no. 11, pp. 1216–1223, 2003.

[38] J. Edeline and N. Weinberger, “Thalamic short-term plasticity in the auditory

system: Associative retuning of receptive fields in the ventral medial geniculate body." *Behavioral Neuroscience*, vol. 105, no. 5, pp. 618–639, 1991.

[39] M. Brosch, E. Selezneva, and H. Scheich, "Nonauditory events of a behavioral procedure activate auditory cortex of highly trained monkey," *The Journal of Neuroscience*, vol. 25, no. 29, pp. 6797–2806, 2005.

[40] A. S. Bregman, *Auditory scene analysis: the perceptual organization of sound.* Cambridge, Mass.: MIT Press, 1990.

[41] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in cognitive sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[42] C. Alain and S. R. Arnott, "Selectively attending to auditory objects," *Frontiers in bioscience : a journal and virtual library*, vol. 5, pp. D202–12, Jan 1 2000.

[43] S. Han and N. Vasconcelos, "Biologically plausible saliency mechanisms improve feedforward object recognition," *Vision Research*, vol. 50, pp. 2295–2307, 2010.

[44] J. T. Serences, S. Yantis, A. Culberson, and E. Awh, "Preparatory activity in visual cortex indexes distractor suppression during covert spatial orienting," *Journal of Neurophysiology*, vol. 92, p. 3538 3545, 2004.

[45] S. Greenberg, "Temporal properties of spoken language," in *Proceedings of the International Congress on Acoustics*, Kyoto, Japan, 2004, pp. 441–445.

[46] T. Kinnunen and H. Lib, “An overview of text-independent speaker recognition: from features to supervectors,” *Speech Commun.*, vol. 52, pp. 12–40, 2010.

[47] C. E. Schreiner, “Order and disorder in auditory cortical maps,” *Current Opinion in Neurobiology*, vol. 5, no. 4, pp. 489–496, 1995.

[48] M. A. Escabi, R. Nassiri, L. M. Miller, C. E. Schreiner, and H. L. Read, “The contribution of spike threshold to acoustic feature selectivity, spike information content, and information throughput,” *Journal of Neuroscience*, vol. 25, no. 41, pp. 9524–9534, 2005.

[49] J. Woojay and B. Juang, “Speech analysis in a model of the central auditory system,” *IEEE Trans.Speech and Audio Process.*, vol. 15, pp. 1802–1817, 2007.

[50] Q. Wu, L. Zhang, and G. Shi, “Robust speech feature extraction based on gabor filtering and tensor factorization,” *Proc.IEEE Int.Conf.Acoust.Sp.Sig.Process.*, 2009.

[51] R. Stern, “Applying physiologically-motivated models of auditory processing to automatic speech recognition,” in *International Symposium on Auditory and Audiological Research*, ser. Speech Perception and Auditory Disorders, 2011.

[52] S. A. Shamma, *Lateral inhibition network*, ser. Methos of Neuronal modeling. MIT Press, 1998, pp. 411–460.

[53] W. Byrne, J. Robinson, and S. Shamma, “The auditory processing and recognition of speech,” in *Proceedings of the speech and natural language workshop*, 1989.

[54] K. Wang and S. A. Shamma, “Self-normalization and noise-robustness in early auditory representations,” *IEEE Trans.Speech and Audio Process.*, vol. 2, pp. 421–435, 1994.

[55] H. Versnel, N. Kowalski, and S. A. Shamma, “Ripple analysis in ferret primary auditory cortex. iii. topographic distribution of ripple response parameters,” *Journal of Auditory Neuroscience*, vol. 1, pp. 271–286, 1995.

[56] T. M. Elliott and F. E. Theunissen, “The modulation transfer function for speech intelligibility,” *PLoS computational biology*, vol. 5, no. 3, p. e1000302, 2009.

[57] 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.

[58] T. Cover and J. Thomas, *Elements of information theory*, 2nd ed. Wiley-Interscience, 2006.

[59] H. Bourlard and N. Morgan, *Connectionist speech recognition: A hybrid approach*. Kluwer Academic, Dordrecht, 1994, p. 348.

[60] E. Trentin and M. Gori, “Robust combination of neural networks and hidden markov models for speech recognition,” *Neural Networks, IEEE Transactions on*, vol. 14, no. 6, pp. 1519–1531, 2003.

[61] A. I. Garcia-Moral, R. Solera-Urena, C. Pelaez-Moreno, and F. D. de Maria, “Data balancing for efficient training of hybrid ann/hmm automatic speech recognition systems,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 468–481, 2011.

[62] K. F. Lee and H. W. Hon, “Speaker-independent phone recognition using hidden markov models,” *IEEE Trans.Acoust., Speech, Signal Process.*, vol. 37, pp. 1641–1648, 1989.

[63] M. D. Richard and R. P. Lippmann, “Neural network classifiers estimate bayesian a posteriori probabilities,” *Neural computation*, vol. 3, no. 4, pp. 461–483, 1991.

[64] J. Pinto, S. V. S. S. Garimella, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, “Analyzing mlp. based hierarchical phoneme posterior probability estimator,” *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, pp. 225–241, 2011.

[65] S. Garimella, S. Nemala, N. Mesgarani, and H. Hermansky, “Data-driven and feedback based spectro-temporal features for speech recognition,” *Signal Processing Letters, IEEE*, vol. 17, no. 11, pp. 957–960, nov. 2010.

[66] “Nist 2008 speaker recognition evaluation,” 2008.

[67] S. Nemala, K. Patil, and M. Elhilali, “Multistream bandpass modulation features for robust speech recognition,” in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 1277–1280.

[68] H. G. Hirsch, “Fant: Filtering and noise adding tool,” *http://dnt.kr.hsnr.de/download.html (date last viewed 11/25/2011)*, 2005.

[69] H. Hermansky and N. Morgan, “Rasta processing of speech,” *IEEE Trans.Speech and Audio Process.*, vol. 2, no. 4, pp. 382–395, 1994.

[70] S. Seneff, “A computational model for the peripheral auditory system: Application of speech recognition research,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’86.*, vol. 11, 1986, pp. 1983–1986.

[71] S. W. Beet and I. R. Gransden, “Interfacing an auditory model to a parameteric speech recogniser,” in *Proceedings of the Institute of Acoustics (IOA)*, vol. 14, 1992, pp. 321–321–328.

[72] O. Ghitza, “Auditory models and human performance in tasks related to speech coding and speech recognition,” *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 115–132, 1994.

[73] C. ying Lee, J. Glass, and O. Ghitza, “An efferent-inspired auditory model front-end for speech recognition,” in *12th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011.

[74] N. R. Clark, G. J. Brown, T. Jurgens, and R. Meddis, “A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise,” *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1535–1541, Sep 2012.

[75] Y. K. Muthusamy, R. A. Cole, and M. Slaney, “Speaker-independent vowel recognition: spectrograms versus cochleagrams,” in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, 1990, pp. 533–536.

[76] R. D. Patterson, T. C. Walters, J. Monaghan, C. Feldbauer, and T. Irino, “Auditory speech processing for scale-shift covariance and its evaluation in automatic speech recognition,” in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 2010, pp. 3813–3816.

[77] G. J. Brown, J. Barker, and D. Wang, “A neural oscillator sound separator for missing data speech recognition,” in *Neural Networks, 2001. Proceedings. IJCNN ’01. International Joint Conference on*, vol. 4, 2001, pp. 2907–2912 vol.4.

[78] J. Barker, N. Ma, A. Coy, and M. Cooke, “Speech fragment decoding tech-

niques for simultaneous speaker identification and speech recognition," *Computer, Speech and Language*, vol. 24, no. 1, pp. 94–111, 1 2010.

[79] M. Fanty, R. Cole, and M. Slaney, "A comparison of dft, plp and cochleagram for alphabet recognition," in *Signals, Systems and Computers, 1991. 1991 Conference Record of the Twenty-Fifth Asilomar Conference on*, 1991, pp. 326–329 vol.1.

[80] C. R. Jankowski and R. P. Lippmann, "Comparison of auditory models for robust speech recognition," in *Proceedings of the workshop on Speech and Natural Language.* Association for Computational Linguistics, 1992, pp. 453–453–454.

[81] D. P. Ellis, H.Hermansky, and S. Sharma, "Tandem connection- ist feature extraction for conventional hmm systems," *Proc.IEEE Int.Conf.Acoustics, Speech, and Signal Process.*, 2000.

[82] S. Handel, *Listening: An introduction to the perception of auditory events.* Cambridge, MA: MIT Press, 1993.

[83] P. T. Ansi, "Psychoacoustical terminology," *New York: American National Standards Institute*, 1973.

[84] H. Helmholtz, *On the Sensations of Tone, Dover, New York.* New York: Dover Publications, 1877.

[85] S. M. J A Burgoyne, “A meta-analysis of timbre perception using nonlinear extensions to clascal,” in *Proceedings of the Computer Music Modeling and Retrieval, 2007*, Copenhagen, Denmark, 2007, pp. 181–202.

[86] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, “The timbre toolbox: extracting audio descriptors from musical signals,” *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[87] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, “Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones,” *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 471–482, 2005.

[88] A. Waibel and K. Lee, *Readings in speech recognition.* Morgan Kaufmann Pub. Inc, 1990.

[89] H. Eidenberger, *Fundamental Media Understanding.* atpress, 2011.

[90] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[91] M. McKinney and J. Breebaart, “Features for audio and music classification,” in *International Symposium on Music Information Retrieval*, vol. 3.2.3, 2003.

[92] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics.* Wiley-IEEE Press, 2012.

[93] P. Herrera-Boyer, G. Peeters, and S. Dubnov, “Automatic classification of musical instrument sounds,” *Journal of New Music Research*, vol. 32, no. 1, pp. 3–21, 2003.

[94] J. J. Burred, M. Haller, S. Jin, A. Samour, and T. Sikora, *Audio Content Analysis Semantic Multimedia and Ontologies.* Springer London, 2008, pp. 123–162.

[95] M. D. Lucia, S. Clarke, and M. M. Murray, “A temporal hierarchy for conspecific vocalization discrimination in humans,” *Journal of Neuroscience*, vol. 30, no. 33, pp. 11 210–11 221, 2010.

[96] N. Staeren, H. Renvall, F. D. Martino, R. Goebel, and E. Formisano, “Sound categories are represented as distributed patterns in the human auditory cortex,” *Current Biology*, vol. 19, no. 6, pp. 498–502, 2009.

[97] E. Formisano, F. D. Martino, M. Bonte, and R. Goebel, “”who” is saying ”what”? brain-based decoding of human voice and speech,” *Science*, vol. 322, no. 5903, pp. 970–973, 2008.

[98] C. A. Atencio and C. E. Schreiner, “Laminar diversity of dynamic sound processing in cat primary auditory cortex,” *Journal of neurophysiology*, vol. 103, no. 1, pp. 192–205, 2010.

[99] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “Rwc music database:

Music genre database and musical instrument sound database," *Proceedings of International Symposium on Music Information Retrieval*, pp. 229–230, 2003.

[100] T. R. Agus, C. Suied, S. J. Thorpe, and D. Pressnitzer, "Fast recognition of musical sounds based on timbre," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4124–4133, 2012.

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

[102] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Differential dynamic plasticity of a1 receptive fields during multiple spectral tasks," *Journal of Neuroscience*, vol. 25, no. 33, pp. 7623–7635, 2005.

[103] ——, "Adaptive changes in cortical receptive fields induced by attention to complex sounds," *Journal of Neurophysiology*, vol. 98, no. 4, pp. 2337–2346, 2007.

[104] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design," *Journal of computational neuroscience*, vol. 9, no. 1, pp. 85–111, 2000.

[105] N. Cristianini and J. Shawe-Taylor, *Introduction to support vector machines and*

other kernel-based learning methods. Cambridge, UK: Cambridge University Press, 2000.

[106] D. L. Donoho *et al.*, “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS Math Challenges Lecture*, pp. 1–32, 2000.

[107] C. E. Schreiner and M. L. Sutter, “Topography of excitatory bandwidth in cat primary auditory cortex: single-neuron versus multiple-neuron recordings,” *Journal of Neurophysiology*, vol. 68, no. 5, pp. 1487–1502, 1992.

[108] C. E. Schreiner, J. Mendelson, M. W. Raggio, M. Brosch, and K. Krueger, “Temporal processing in cat primary auditory cortex,” *Acta Oto-Laryngologica*, vol. 532, pp. 54–60, 1997.

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

[110] M. Elhilali, J. B. Fritz, D. J. Klein, J. Z. Simon, and S. A. Shamma, “Dynamics of precise spike timing in primary auditory cortex,” *Journal of Neuroscience*, vol. 24, no. 5, pp. 1159–1172, 2004.

[111] G. B. Christianson, M. Sahani, and J. F. Linden, “The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields,” *The Journal of neuroscience*, vol. 28, no. 2, pp. 446–455, 2008.

[112] S. V. David, N. Mesgarani, J. B. Fritz, and S. A. Shamma, “Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli,” *Journal of Neuroscience*, vol. 29, no. 11, pp. 3374–3386, 2009.

[113] S. Sadagopan and X. Wang, “Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex,” *Journal of Neuroscience*, vol. 29, no. 36, pp. 11 192–11 202, 2009.

[114] M. Elhilali, J. B. Fritz, T. S. Chi, and S. A. Shamma, “Auditory cortical receptive fields: stable entities with plastic abilities,” *Journal of Neuroscience*, vol. 27, no. 39, pp. 10 372–10 382, 2007.

[115] A. Livshin and X.Rodet, “Musical instrument identification in continuous recordings,” in *Conference on Digital Audio Effects*, Naples, Italy, 2004.

[116] J. J. Burred, A. Robel, and T. Sikora, “Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds,” *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 3, pp. 663–674, 2010.

[117] A. G. Krishna and T. V. Sreenivas, “Music instrument recognition: from isolated notes to solo phrases,” in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP ’04). IEEE International Conference on*, vol. 4, 2004, p. iv.

[118] J. Marques and P. J. Moreno, “A study of musical instrument classification using gaussian mixture models and support vector machines,” Compaq Corporation, Cambridge Research laboratory, Tech. Rep., 1999.

[119] J. C. Brown, O. Houix, and S. McAdams, “Feature dependence in the automatic identification of musical woodwind instruments,” *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1064–1072, 2001.

[120] T. Kitahara, M. Goto, and H. G. Okuno, “Musical instrument identification based on f0-dependent multivariate normal distribution,” in *in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Hong Kong*, 2003, pp. 409–412.

[121] A. Eronen and A. Klapuri, “Musical instrument recognition using cepstral coefficients and temporal features,” *Proceedings of the 2000 IEEE Conference on Acoustics, Speech, and Signal Processing.*, vol. 2, pp. II753–II756, 2000.

[122] G. Agostini, M. Longari, and E. Pollastri, “Musical instrument timbres classification with spectral features,” in *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, 2001, pp. 97–102.

[123] A. Livshin and X. Rodet, “The significance of the non-harmonic noise versus the harmonic series for musical instrument recognition,” in *in Proc. of the 7th International Conference on Music Information Retrieval (ISMIR*, 2006, pp. 95–100.

[124] B. Kostek, “Musical instrument classification and duet analysis employing music information retrieval techniques,” *Proceedings of the IEEE*, vol. 92, no. 4, pp. 712–729, 2004.

[125] J. Marozeau, A. de Cheveigne, S. McAdams, and S. Winsberg, “The dependency of timbre on fundamental frequency,” *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2946–2957, 2003.

[126] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling.* Chapman and Hall, 2001.

[127] S. Dixon, “Onset detection revisited,” in *Proc. of the 9th Int. Conference on Digital Audio Effects*, Montreal, Canada, Sep. 18–20, 2006.

[128] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, “A tutorial on onset detection in music signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[129] J. L. Goldstein, “An optimum processor theory for the central formation of the pitch of complex tones,” *Jounal of the Acoustical Society of America*, vol. 54, pp. 1496–1516, 1973.

[130] J. Yang, R. Yan, and A. G. Hauptmann, “Cross-domain video concept detection using adaptive svms,” in *MM’07*, 2007.

[131] O. Joly, F. Ramus, D. Pressnitzer, W. Vanduffel, and G. A. Orban,

"Interhemispheric differences in auditory processing revealed by fmri in awake rhesus monkeys," *Cerebral Cortex*, vol. 22, no. 4, pp. 838–853, 2011.

[132] T. D. Griffiths and J. D. Warren, "What is an auditory object?" *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 887–892, 2004.

[133] C. M. Karns and R. T. Knight, "Intermodal auditory, visual, and tactile attention modulates early stages of neural processing," *Journal of cognitive neuroscience*, vol. 21, no. 4, pp. 669–683, Apr 2009.

[134] V. Poghosyan and A. A. Ioannides, "Attention modulates earliest responses in the primary auditory and visual cortices," *Neuron*, vol. 58, no. 5, pp. 802–813, Jun 12 2008.

[135] L. Whiteley and M. Sahani, "Attention in a bayesian framework," *Frontiers in Human Neuroscience*, vol. 6, 2012.

[136] J. B. Fritz, S. V. David, D. Winkowski, P. Y. n, M. Elhilali, and S. A. Shamma, "Intention and attention: Top-down influences on the representation of task-relevant sounds," 2010 2010.

[137] K. Krumbholz, S. B. Eickhoff, and G. R. Fink, "Feature- and object-based attentional modulation in the human auditory "where" pathway," *Journal of cognitive neuroscience*, vol. 19, no. 10, pp. 1721–1733, Oct 2007.

[138] *The BBC Sound Effects Library Original Series.* http://www.sound-ideas.com, May 2006.

[139] J. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain House, New Paltz, NY*, October 2013.

[140] F. Eyben, M. Wllmer, and B. Schuller, "opensmile: The munich versatile and fast open-source audio feature extractor," in *ACM Multimedia (MM), Florence, Italy*, 2010, p. 14591462.

[141] M. Barnard and J.-M. Odobez, "Robust playfield segmentation using map adaptation," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, aug. 2004, pp. 610 – 613 Vol.3.

[142] S. Marcel and J. Millan, "Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 743 –752, april 2007.

[143] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[144] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: the

biological bases of musical timbre perception," *PLoS computational biology*, vol. 8, no. 11, p. e1002759, Nov 2012.

[145] K. Patil and M. Elhilali, "Task-driven attentional mechanisms for auditory scene recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. accepted, 2013 2013.

[146] S. K. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 416–426, 2013.

[147] S. A. Shamma and D. J. Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2631–2644, 2000.

Vita

Kailash Patil was born in 1986 in a small town called Basavakalyan in south India. He received primary education in Bangalore. He was admitted to Indian Institute of Technology Guwahati in 2004 where he received the B. Tech. degree in Electronics and Computer Engineering in 2008. He then enrolled in the Electrical and Computer Engineering(ECE) Ph.D. program at Johns Hopkins University in 2008. During the course of his doctoral studies he received M.S.E degree from ECE department in 2011.

Starting in November 2013, he will be joining Pindrop Security as an Audio Research Scientist.