A HIERARCHY BASED ACOUSTIC FRAMEWORK FOR AUDITORY SCENE ANALYSIS

by

Debmalya Chakrabarty

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

April, 2018

Abstract

The acoustic environment surrounding us is extremely dynamic and unstructured in nature. Humans exhibit a great ability at navigating these complex acoustic environments, and can parse a complex acoustic scene into its perceptually meaningful objects, referred to as “auditory scene analysis". Current neuro-computational strategies developed for auditory scene analysis related tasks are primarily based on prior knowledge of acoustic environment and hence, fail to match human performance under realistic settings, i.e. the acoustic environment being dynamic in nature and presence of multiple competing auditory objects in the same scene. In this thesis, we explore hierarchy based computational frameworks that not only solve different auditory scene analysis related paradigms but also explain the processes driving these paradigms from physiological, psychophysical and computational viewpoint.

In the first part of the thesis, we explore computational strategies that can extract varying degree of details from complex acoustic scene with an aim to capture non-trivial commonalities within a sound class as well as differences across sound classes. We specifically demonstrate that a rich feature space of spectro-temporal modulation representation complimented with markovian based temporal dynamics information captures the fine and subtle

changes in the spectral and temporal structure of sound events in a complex and dynamic acoustic environment. We further extend this computational model to incorporate a biologically plausible network capable of learning a rich hierarchy of localized spectro-temporal bases and their corresponding long term temporal regularities from natural soundscape in a data driven fashion. We demonstrate that the unsupervised nature of the network yields physiologically and perceptually meaningful tuning functions that drive the organization of acoustic scene into distinct auditory objects.

Next, we explore computational models based on hierarchical acoustic representation in the context of bottom-up salient event detection. We demonstrate that a rich hierarchy of *local* and *global* cues capture the salient details upon which the bottom-up saliency mechanisms operate to make a "new" event pop out in a complex acoustic scene. We further show that a top-down event specific knowledge gathered by scene classification framework biases bottom-up computational resources towards events of "interest" rather than any new event. We further extend the top-down framework in the context of modeling a broad and heterogeneous acoustic class. We demonstrate that when an acoustic scene comprises of multiple events, modeling the global details in the hierarchy as a mixture of temporal trajectories help to capture its semantic categorization and provide a detailed understanding of the scene.

Overall, the results of this thesis improve our understanding of how a rich hierarchy of acoustic representation drives various auditory scene analysis paradigms and how to integrate multiple theories of scene analysis into a unified strategy, hence providing a platform for further development of computational scene analysis research.

Thesis Committee

Primary Readers

Dr. Mounya Elhilali (Primary Advisor)
Associate Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dr. James West
Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Alternate Reader

Dr. Shinji Watanabe
Associate Research Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Acknowledgments

I want to first express my sincere thanks and gratitude to my advisor, Dr. Mounya Elhilali for her continued patience, guidance, mentorship, understanding, and support throughout my studies and for allowing me to explore different areas and ideas in the field of auditory research over the years. I also want to thank my committee members, Dr. James West and Dr. Shinji Watanabe for taking the time to serve on my dissertation committee. I also want to thank Professor Hynek Hermansky for motivating me to apply to Johns Hopkins University and Dr. Sanjeev Khudanpur for helping me out with internship opportunity at Xerox Research.

I want to thank my friends in the Laboratory for Computational Audio Perception: Kailash, Mike, Merve, Ashwin, Ben, Dimitra, Nick, David, Harsha and Sandeep. You were all a pleasure to work with and thanks for all the meaningful and productive discussions which helped me in my research from time to time. I also want to thank all my friends in JHU, especially: Suman Da, Sayantan Da, Arijit Da, Sayak, Anindya, Iqbal and Nagal. Thanks for making JHU such a wonderful and jovial place and giving me such beautiful memories to be cherised forever.

Last but not the least, I would like to thank my parents (Ma and Babai) for

their continued love, patience and support. This journey wouldn't have been possible without their continued encouragement throughout different phases of my life at JHU. Finally, I would like to thank my elder brother (Bhaiya) , who always used to believe in my abilities and continues to inspire me with his simplicity and honesty.

Table of Contents

List of Tables

List of Figures

Abbreviations

ASA	Auditory Scene Analysis
A1	Auditory Cortex
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
K-NN	K-Nearest Neighbor
BN	Bayesian Network
DNN	Deep Neural Network
LSTM	Long Short Term Memory
TSVD	Tensor Singular Value Decomposition
MTT	Mixture of Temporal Trajectories
STRF	Spectro-Temporal Receptive Field
MFCC	Mel Frequency Cepstral Coefficients
RBM	Restricted Boltzmann Machine
CRBM	Conditional Restricted Boltzmann Machine
mCRBM	Mixture of Conditional Restricted Boltzmann Machine
CD	Contrastive Divergence
MTF	Modulation Transfer Function

Notation

GF	Gabor Filter
$\mathcal{R}$	Modulation features
$S(t, f)$	Auditory spectrogram
$\mathcal{L}_1$	RBM layer
$\mathcal{L}_2$	CRBM layer
r_k	RBM Filter response
o_k	CRBM Filter response
τ	CRBM rate of Temporal Integration
W	RBM interconnection weights
θ	CRBM parameter set
V	Synaptic weight matrix
Ω	Spectral modulation
ω	Temporal modulation
H	Harmonicity neurons
O	Onset neurons
S	Slow neurons
F	Fast neurons
ΔF	Frequency difference
ΔAM	Amplitude modulation difference
g	Adaptation gain
$\hat{y}_t$	Class Posterior probability
$C_{abnormal}$	Set of abnormal events
X	Training set

Chapter 1

Introduction

1.1 Background and Motivation

In everyday life, humans are surrounded by multitude of acoustic scenes e.g. a cocktail party, a busy street, or a crowded coffee shop where the sound originating from a particular source do not exist in isolation. They persistently occur in presence of other competing sources and distractors that form a person's acoustic environment. The auditory system performs many intricate steps involved in processing of sound which enables our ability to hear and perceive constantly changing acoustic environment. This process of auditory system to parse the complex mixture such that the entire soundscape gets organized into meaningful 'percepts' is referred to as 'auditory scene analysis' (ASA) [1].

Humans and animals have got great ability to interpret and understand the complex acoustic environment surrounding them. For instance, following a conversation at a noise cocktail party or locating your newborn in the middle of a colony of screaming birds relies on successfully segregating sounds into

meaningful streams. However, the neural mechanisms by which the human brain achieves this feat are largely unknown. In last three decades, auditory scientists have postulated several theories for explaining the processing of complex auditory scenes and perceptual phenomenon driving scene segregation paradigm. Many of the computational models have been developed based on these theories. However, the diversity of all these models and continued interest in this field suggests that auditory research community still hasn't solved the problem of complex scene segregation from computational perspective. Most of the existing computational models are still not able to match human performance when faced with realistic setting, such as, variability of noise environment and presence of competing sound sources among others.

This thesis concerns computational models of auditory scene analysis, with a focus on how sound is represented across different stages of auditory pathway and how does this representation gets transformed into an auditory 'percept'. In particular, we consider models that seek to explain: (1) how to represent different stages of processing of sound in auditory system via efficient computational strategies (2) what are the different degrees of transformation that the sound goes through across each stage of processing. (3) how do the details extracted in each stage finally combine to form an auditory percept. (4) how to adapt the representation captured at each stage to constantly changing acoustic environment. In the course of this thesis, we show that the proposed computational models not only succeeds in capturing efficient representation of sound but also successfully drives many of the paradigms related to auditory scene analysis.

1.2 Key Stages of Auditory Scene Analysis

Based on past theories, ASA has been conceptualized as a multi-stage process as shown in Figure 1.1. In the first stage, the acoustic mixture is decomposed

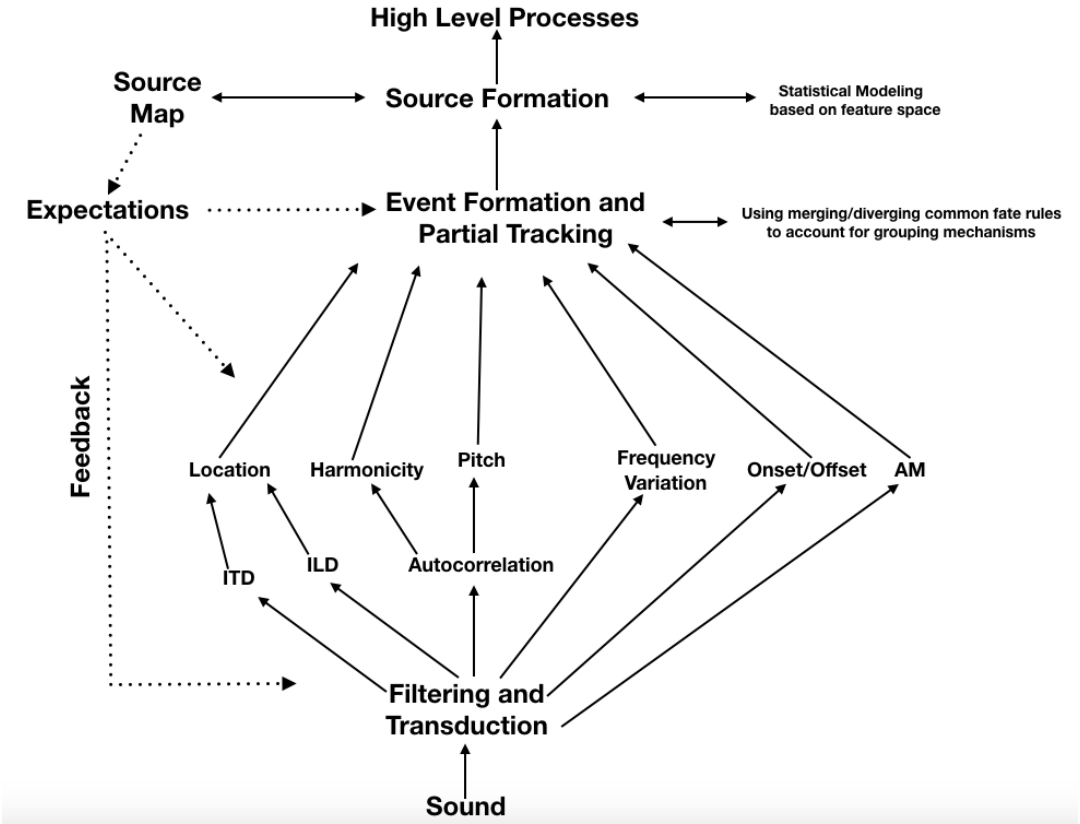

Figure 1.1: Block diagram of stages of Auditory Scene Analysis (adapted from Mellinger,1992). Event and source formation are driven by feature representation and prior knowledge

into feature elements. These feature elements are considered to be the building blocks of auditory scene, which describes different acoustic events. Since past

many years, a lot of effort is being thrusted upon obtaining a feature representation that is capable of extracting relevant informations from complex auditory scenes. This feature representation is complimented with a grouping mechanism that combines the building blocks arising from the same acoustic source, hence, finally forming a perceptual structure of the source called a *stream*. These grouping mechanisms are well guided by feature dynamics characterizing the constantly changing objects within the scene [2]. Slower dynamics in acoustic signals are believed to be the main carrier of information in complex acoustic scenes [3] as well as commensurate with temporal dynamics of stream formation and auditory grouping [4]. The second stage comprises of statistical modeling based approach which projects the groups of feature representation onto a space of auditory objects via learning some useful statistics demonstrative of key details present in an acoustic scene. These statistical models typically form the core of high level processes and are based on past knowledge and expectations. Hence, the ability of a computational model in any auditory scene analysis related paradigm depends on two key components : 1) Obtaining a rich and robust feature representation that can capture object specific details present in the scene and compliment it with grouping mechanisms matching their temporal regularities to dynamics of an auditory object 2) Using the feature space to build statistical model based representation of the auditory object to drive the key processes of auditory scene analysis.

1.2.1 Feature Representation and Grouping Mechanisms

It is well argued in vision literature that a high dimensional feature representation can empower a computational visual recognizer in extracting meaningful semantics from arbitrary natural images [5]. Along similar lines, kernel based approaches have proved to be very successful in visual recognition or classification based tasks [6]. In recent times, unsupervised deep layered architectures have been proved to be extremely effective in capturing the salient aspects of a visual scene through discovery of discriminative clusters, parts, mid-level features and hidden units [7]. In recent past, a number of these deep network techniques have been extended to time series visual data, thereby modeling temporal trajectories of dynamic visual scene and apply it to human pose estimation and motion tracking problems [8]. It has also been well established in [9] that visual cues, even when well distributed in space bind over time to form a perceptual and unified organization of visual objects [10]. The “scene segmentation" problem in vision is analogous to the problem of auditory scene analysis since past psychoacoustic studies have shown that the auditory system also uses a combination of rich feature space and temporal grouping strategies to process an auditory scene [11, 12].

The concept of a rich feature representation and temporal grouping mechanisms has been well observed in past physiological studies. There are physiological evidences which suggest that the auditory pathway in human brain performs the task of decomposing the acoustic signal into its constituting elements and mapping them into perceptual streams [13]. The initial transformation of acoustic signal into set of frequency components along the tonotopic

axis happens at the cochlea [14]. As the signal pass onwards from the cochlea towards the auditory cortex (A1), additional features are extracted forming a rich feature representation that includes onsets, offsets, harmonicities, amplitude and frequency modulations (AM, FM). These features form the backbone of a sound object representation in human auditory system [15–17]. Recent neurophysiological studies have also suggested that a rich high dimensional feature representation doesn't provide the complete picture of how sound is represented in human brain, hence emphasizing the importance of grouping mechanisms like temporal coherence theory [18]. These experiments suggests that a combination of rich feature representation and feature binding induced by temporal coherence provides a more complete and robust representation of sound along the auditory pathway and facilitates segregation of auditory objects into perceptual streams.

An important aspect of ASA based computational models is obtaining a computational medium for parameterizing the auditory feature space [23]. Robust low-level acoustic features have been proven to be effective representations for variety of auditory scene analysis related paradigms [19–22]. In addition, Mel Frequency Cepstral Coefficients (MFCC) have been a popular choice of acoustic scene classification as they are quite powerful in capturing the overall 'transfer function' of a scene [23, 24]. However, for high-level interpretation of a complex scene, one has to rely on subtle but intricate details which low-level features fail to capture. Instead, it is imperative to consider signal features that capture the spectral and temporal modulations (i.e. changes) in the scene over a wide range of resolutions [25]. Gabor features offer such flexibility in time and frequency by tracking the localized spectral and temporal signal

changes over various scales [26]. In recent times, a very common technique of characterizing an auditory feature space is by projecting time-frequency spectrogram representation of sound onto a two dimensional filter; typically referred to as spectro-temporal receptive fields (STRF) [27–30]. The primary assumption in these computational frameworks is that STRFs are linear filters that map the raw acoustic signal to multidimensional feature in a linear transformation process; however a lot of physiological studies suggest that neural representation of sound in auditory pathway is a nonlinear process characterized by nonlinearity of A1 responses [31]. Another major shortcoming of these models is parameterized transformation of acoustic signal to feature space. Such a transformation gets limited in its scope of capturing a feature space that can encode any auditory soundscape irrespective of its complexity [32].

1.2.2 Statistical modeling for Auditory Scene Analysis

Majority of statistical based approaches based on biological mechanisms for scene analysis processes are evaluated in terms of how faithfully they are able to reproduce the psychoacoustics of stream segregation. For example, Wang et. al in [33] measure the fitness of their model based on its ability to reproduce the fission and temporal coherence boundaries reported in [34]. Some more examples based in this approach include models of auditory periphery explaining the streaming paradigm of two tone sequences [35, 36], stream segregation based on pitch modulations [37], as well as some models based on elaborate sound sequences like segregating speech from environmental

sounds (e.g siren, telephone rings etc.) [38]. These models are restricted to the analysis of how a specific cue contributes to stream segregation processes. These models also lack in giving an insight about what is the underlying representation that captures these wide range of psychoacoustic cues from input sound as suggested by physiological studies [13]. This suggests that there is an existing gap between the computational models based on psychoacoustic theories and physiological bases of how sound is represented in human brain. Some of the computational models also explore the idea of temporal coherence from stream segregation perspective and are primarily based on the notion that acoustic features of sounds emitted by the same source recur together. Hence, grouping by temporal coherence bind them together into single auditory object representation [18, 39, 40]. These models are typically built on the framework of capturing long term temporal correlation across multi-dimensional feature representation assumed to provide a good mathematical solution supporting the theory of temporal coherence. However, the backbone of these models is primarily based on prior knowledge of what sources being present in the scene. These models are typically driven by supervised training framework learning the mapping of sound representation from feature space to an auditory object. Hence, it can be well argued that these models lack the biological perspective of temporal grouping mechanisms [41] driving the formation of auditory objects based on feature association.

Over past several decades, many research fields related to auditory scene analysis have been studied to varying extent. The related research comprises different audio classification problems such as speech/music discrimination [42], sound source classification [43], environmental noise classification [44],

content-based audio classification [45], scene segregation [46] and salient event detection among others [47]. In the context of scene analysis, scene classification typically refers to the task of labeling an acoustic scene with its corresponding *class*. However, *event detection* task refers to detection of onset/offset for the classes present in a recording and classification within the estimated onset/offset, which is typically a requirement in real-life scenario. The complexity of such tasks typically lie in the unstructured nature of sounds events and high degree of variation present among the sound events from the same class [48]. Moreover, there can be multiple sound sources that produce sound events belonging to the same class, e.g. a household ambience typically comprises of sound events like water gushing out of a tap, sound of cooking utensils etc. which adds to the semantic richness of sound class and hence increases the complexity of underlying task.

In past efforts, auditory scientists have attempted to solve these problems using statistical modeling based approaches. For example, several techniques for auditory scene classification have been based on descriptive statistics quantifying various aspects of statistical distributions including moments (such as mean, variance, skewness and kurtosis of distribution) derived from feature representation of acoustic scenes [49]. Other methods employed in scene classification and event detection task include generative architectures like Gaussian Mixture Models (GMMs) in which features vectors are interpreted as being generated by multi-modal distribution expressed as sum of Gaussian distributions [22, 50]. The other class of generative models used in several classification and detection tasks are based on Hidden Markov

Models (HMMs) to account for the temporal unfolding of events within complex soundscapes [51]. Source separation and scene segregation related tasks have been primarily based on certain segmentation rules based on perceptual grouping cues [52] semi-manually designed to operate on low-level features to estimate a time-frequency mask that isolates the signal components belonging to different sources [53]. Non-negative matrix factorization (NMF) [54–56] has been another popular technique aimed to learn a set of non-negative bases representation used to estimate mixing factors. Statistical models like support vector machines (SVM) and Bayesian Network (BN) have also been employed to learn relationship between audio effects and high-level scene representations [57].

In recent times, with the emergence of deep learning architectures, deep learning methods have proved to be extremely successful and have demonstrated state-of-the art results in vision related tasks like visual scene segmentation, classification and identification among others [58–60]. Taking cue from vision based studies, auditory scientists have applied such techniques to areas of speech recognition, speaker identification etc. and have demonstrated state of the art performance [61, 62]. However, the applicability of such techniques have also been extended to tasks related to auditory scene analysis as well as mentioned above. Deep learning models like Feed Forward Neural Networks (FNNs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been shown to perform significantly better compared to well established statistical model based approaches like GMM-HMM, SVM and NMF for acoustic scene classification [63], monophonic/polyphonic event detection [64] and source separation [65] among others. Recent efforts have

also been employed in the direction of learning rich feature space from natural soundscape in a data driven fashion, and subsequently using these spaces in domains like music genre classification, phoneme classification and speaker identification [66, 67].

With all the advancement of computational strategies in tackling the problem of auditory scene analysis, there still lacks a proper study which presents a comprehensive overview of different theories of scene analysis and tie them up into a common computational framework. As suggested by Bregman in [1], auditory scene analysis paradigm is driven by two types of perceptual grouping theories namely *primitive grouping* and *schematic grouping*. Primitive grouping is driven by incoming acoustic signal and is primarily described by Gestalt principles of perceptual organization [68]. In contrast, schematic grouping employs the knowledge of familiar patterns and concepts that have been acquired through experience of acoustic environments.There have been a number of physiological and psychoacoustic studies done to corroborate these theories in context of stream segregation [69, 70]. However, most of the neuro-computational strategies of auditory scene explored in the literature rely on object based auditory supported by backend classifiers for building a scene analysis framework. In this thesis, we aim to present a comprehensive analysis of both primitive and schematic grouping principles by integrating them into a common computational framework and show how such a framework can drive different scene analysis paradigms like scene segregation, classification and detection among others.

1.3 Thesis Contributions and Outline

In this section, we present the outline of the thesis and state some of the main contributions of the thesis. In the introduction, we provided some of the background and motivations for research in the topic of auditory scene analysis. Following the introduction in Chapter 2, we present a framework that explores the role of temporal dynamics in a task of acoustic scene classification. The framework presented in this chapter is primarily driven by features generated from a cortical model to extract information in spectral and temporal modulations. In this paper, we propose a framework that provides a detailed local analysis of spectro-temporal modulations augmented with generative modeling that map both the average modulation statistics of the scene using GMM as well temporal trajectories of these modulations using HMM. One novel contribution of this framework is its ability of to capture the heterogeneity of an unstructured scene like “household ambience" via modeling the temporal dynamics of modulation features, which has not been accounted for in great detail by most of the existing computational strategies. Our analysis shows that a hybrid system of average modulation statistics combined with temporal dynamics information can capture the non-trivial commonalities within a sound class and differences between sound classes.
In Chapter 3, we propose an architecture so as to answer some key questions faced by the framework in Chapter 2. One major question is the spectro-temporal modulation filters being hand-crafted in nature are limited in their scope to specific mapping in modulation space. Another major constraint is the imposition of linearity in their design which restricts the filters in spanning

those acoustic events characterized by non-linear spectro-temporal modulations. In order to seek answers to these questions in this chapter, we propose a computational framework that leverages ideas from different theories of scene organization (both vision and audition) and integrate them into a unified architecture. The proposed architecture learns a rich hierarchy of localized and global "cues" from natural scenes in an unsupervised and non-linear data driven fashion and is well supported by hebbian based grouping mechanism which binds temporally correlated cues together to form perceptual representation of auditory objects in the scene. A major contribution of this work is that the proposed framework successfully replicates well established physiological and psychoacoustic bases of scene segregation across varied complexity of sounds and also quantifies the complimentary role of segregation and binding cues in driving scene segregation processes.

In Chapter 4, we propose an extension to the framework developed in Chapter 3 with an aim to incorporate bottom-up saliency mechanism in the hierarchical architecture. We use the proposed architecture in an abnormal event detection task so as to explore how saliency mechanisms drive event detection paradigm and add to the present literature which lack relevant studies in this context. The proposed bottom-up architecture learns a rich hierarchy of localized and global "attributes" from natural scenes in an unsupervised data driven fashion and is well supported by an adaptation and reset mechanism biasing the "attributes" towards salient details of an event in a complex acoustic scene. A top down acoustic scene classification framework is also developed by modeling the sequential representation of local and global attributes via a deep long short-term memory network (LSTM). We further exploit the complementarity

of bottom-up and topdown processes in context of saliency mechanisms by applying them over abnormal event detection paradigm. A key contribution of this work is that the proposed framework outperforms the individual bottom-up and topdown approach as well as achieves satisfactory performance compared to baseline bottom-up and top-down saliency models in salient event detection task.

Chapter 5 deals with exploring a rich hierarchical characterization of a broad and complex sound class that is capable of capturing nontrivial commonalities within a sound class as well differences across sound classes, with an aim to identify abnormal sound events in a complex acoustic scene. In this chapter, we have proposed a methodology for representing sound classes using a hierarchical network of convolutional features and mixture of temporal trajectories (MTT). The framework couples unsupervised and supervised learning and provides a robust scheme for detection of abnormal sound events in a subway station. The key contribution of this work is the ability of the proposed representation in capturing non-trivial commonalities within a single sound class and variabilities across different sound classes as well as high degree of robustness in noise.

Finally, in Chapter 6, we conclude with a summary of the main contributions of this thesis. We also briefly discuss the possibilities for further research in the area of hierarchy driven acoustic representation and how to develop top-down mechanisms to adapt such representation to constantly changing acoustic environment.

Chapter 2

Role of Temporal Dynamics in Acoustic Scene Classification

2.1 Introduction

Our surrounding soundscapes are constantly changing as we go about our lives; walking from an office to the street to a cafe and carrying conversations along the way. Humans exhibit a great ability at navigating these complex acoustic environments, and can effortlessly parse and identify their acoustic surroundings; in a process called *auditory scene analysis* [1]. This phenomenon describes complex neural and cognitive processes that underly our ability to detect, identify and classify sound objects in complex acoustic environments. Much like one can identify different visual scenes by the attributes of their constituting objects, a similar process takes place allowing our brain to distinguish a human voice from a bird chirp or a car horn [42]. This ability can provide a great degree of robustness and flexibility to technologies like communication aids for sensory-impaired, surveillance and security systems, context aware computing, audio annotation etc.

The 'identity' of an acoustic scene is largely determined by the acoustic characteristics of the sound sources present at the scene. These sources adapt the spectral profile of the signal to reflect the shape and structure of the vibrating bodies, along with trajectories and reflection paths traveled by sounds until they reach the listener's ear or recording device. The analysis of these characteristics for purposes of automatic identification or classification of acoustic scenes has to take into account all the spectral and temporal attributes of the signal. It has to also be sensitive enough to the natural variability in each class of scenes while discriminative enough across classes. A number of scene classification studies have explored the relevance of low-level features in capturing scene characteristics. These features include low-level time based and frequency based descriptors like short-time energy (STE), zero-crossing rate (ZCR), voicing features like periodicity and pitch information, linear predictive coding coefficients (LPC), as well as the energy distribution entropy of discrete Fourier transform components [19, 20, 22, 71]. These reports suggest that low-level acoustic features are powerful in distinguishing simple scenes. In addition, Mel Frequency Cepstral Coefficients (MFCC) have been a popular feature of choice in studies of acoustic scene classification as they are quite powerful in capturing the overall 'transfer function' (or spectral shaping function) of each scene, and have indeed led to a number of successful implementations of event classification systems [23, 24]. However, in case of complex acoustic scenes, the intricate details of the spectral profile and temporal dynamics of sound events in a scene makes applicability of average features rather limited. Use of global representations of a scene such as cepstral coefficients are generally not capable of capturing fine and subtle

changes in the spectrum as it evolves over time; especially in case of dynamic and nonstationary scenes. Instead, it is imperative to consider signal features that capture the spectral and temporal modulations (i.e. changes) in the scene over a wide range of resolutions. Gabor features offer such flexibility in time and frequency by tracking the localized spectral and temporal signal changes over various scales [72].

Use of representative features is intricately linked with choice of backend classifiers that are flexible enough to capture variability across scene classes yet stable enough to work with nuances emerging from the signal features. Commonly used learning techniques include K-NN classifiers and Gaussian mixture models (GMM) which have been used to classify auditory scenes into predefined semantic categories [22]. Statistical models like support vector machines (SVM) and Bayesian network (BN) have also been employed to learn the relationships between audio effects and high-level scene representations [57, 73]. Additional techniques employ descriptive statistics of low level acoustic features and quantify their statistical distributions in terms of mean, variance, skewness and kurtosis [19, 49, 74]. More recently, researchers have focused on modeling the mean statistics obtained from spectro-temporal modulation features via discriminative classifier using multilayer perceptrons and have shown that these representations greatly outperform low-level features like MFCC and its statistics in auditory scene classification task [75].

That being said, one of the challenges of the scene classification task is the inherent complexity of describing what a 'scene' is and the degree granularity that is defined with a chosen dataset for analysis. In the commonly used BBC sound effects dataset [76], the sound class labeled *humor* is a more generic class

that encompasses instances of individuals cheering or laughing. These two 'subclasses' can be rather heterogeneous in their signal characteristics making the use of average feature profiles rather limited. Instead, it appears that combining a rich representation of spectro-temporal changes in the signal along with their temporal trajectories could provide added flexibility to capture the heterogeneity of the audio samples in each class [77]. In the current work, we explore the use of a hybrid systems that combines use of spectro- temporal modulation features along with their temporal dynamics to represent sound classes. We explore use of temporal trajectories beyond the classical derivative parameters (Δ, $\Delta\Delta$) by using Hidden Markov Modeling (HMM) applied to modulation features.

Figure 2.1 shows a schematic representation of proposed architecture. Each of the stages in Figure 2.1 are explained in detail in following sections. The organization of this chapter is as follows: In section 2.2, a brief description of the spectro-temporal modulation features used in the proposed system is provided. Section 2.3 outlines the classification system modeling both mean statistics as well as temporal trajectories of the modulation features. Section 2.4 describes the experimental set up and scene classification results; while section 2.5 provides conclusions and discussion of the results.

2.2 Modulation Based Features

The analysis of modulation features in the acoustic signal is performed in two stages. First, a time-frequency auditory spectrogram is extracted based on a model of peripheral processing in the mammalian auditory system [29]. This

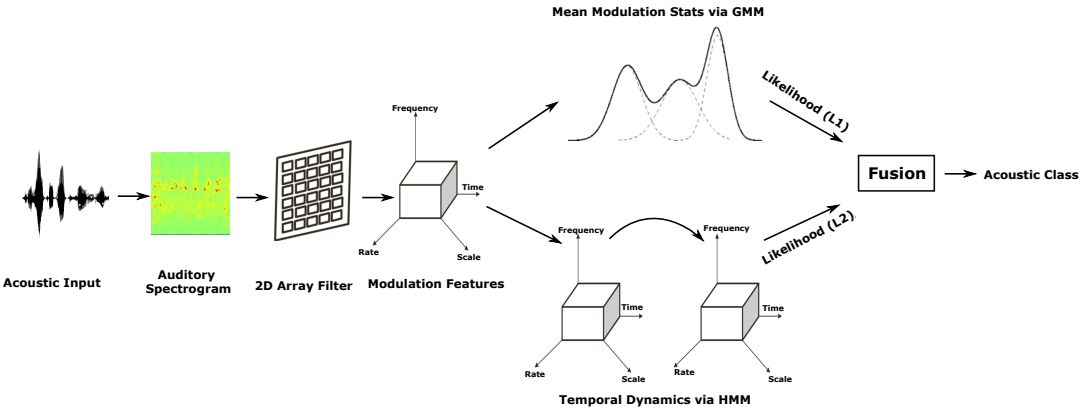

Figure 2.1: A block diagram representation of GMM-HMM based acoustic scene classification framework

first stage starts with a bank of 128 asymmetric filters equally-spaced on a logarithmic axis over 5.3 octaves spanning the range 180 Hz to 8000 Hz. Next, the signal undergoes spectral sharpening via first order derivative along the frequency axis followed by half wave rectification and short term integration with $u(t,\tau) = e^{-t/\tau}u(t)$ where $\tau = 2$ ms. This filterbank analysis results in a time-frequency auditory spectrogram represented by $y(t,f)$. The second stage follows to extract modulation features in the signal. This analysis is performed using a bank of two-dimensional Gabor Filters (GF). Each Gabor filter $GF(f,t;\mathbf{s},\mathbf{r})$ is parameterized by its spectral modulation tuning or scale ($\mathbf{s}$ in cycles/octave) and temporal modulation tuning or rate ($\mathbf{r}$ in Hertz). It effectively filters the detailed fluctuations (called modulations) in the spectral and temporal patterns of the signal. The analysis yields a four-dimensional tensor $\mathcal{R}$ parameterized by time $\mathbf{t}$, frequency $\mathbf{f}$, scale $\mathbf{s}$ and rate $\mathbf{r}$ represented as:

$$\mathcal{R}(t,f;\mathbf{s},\mathbf{r}) = |y(t,f) \otimes_{f,t} GF(f,t;\mathbf{s},\mathbf{r})| \tag{2.1}$$

where $\otimes_{f,t}$ denote convolution in time and frequency. The tensor $\mathcal{R}$ is a multi-resolution mapping of the acoustic signal onto a high-dimensional space [78]. This mapping is akin of the rich representation of sounds in the central mammalian auditory system where specro-temporal response fields of cortical neurons [79] can be mapped onto a space tiled by these Gabor filters.

2.3 Classification of Modulation Features

We use the modulation features denoted by $\mathcal{R}$ to build statistical models for the scene classification task. Our analysis contrasts two types of models, as described next:

2.3.1 Modeling mean statistics of Spectro-Temporal Representation

The first approach builds a generative model of the data in each class based on average statistics of the scenes. Average statistics are obtained from the 4D modulation tensors $\mathcal{R}$ by first integrating the features over the duration of audio segment. For all analyses presented here, we segment recordings of all sound classes over non-overlapping 1s windows. For each segment, we get a mean representation along frequency, rate and scale axes denoted by $\bar{\mathcal{R}}(f, \mathbf{s}, \mathbf{r})$ which can be expressed as:

$$\bar{\mathcal{R}}(f, \mathbf{s}, \mathbf{r}) = E[\mathcal{R}(t, f; \mathbf{s}, \mathbf{r})] \tag{2.2}$$

The tensor $\bar{\mathcal{R}}$ is further projected onto a lower dimensional space using Tensor Singular Value Decomposition (TSVD) [80]. We keep 420 dimensions that

maintain 99 % variance in the data; resulting in a lower-dimensional modulation tensor $\bar{\mathcal{R}}$. Given the use of modulation features over time and frequency, this lower dimensional $\bar{\mathcal{R}}(f, \mathbf{s}, \mathbf{r})$ captures average changes in the audio segment and is used as feature vector to build Gaussian Mixture Models (GMM) [81] of each sound class to learn its inherent statistical characteristics.

2.3.2 Modeling the temporal trajectories of Spectro-Temporal Representation

Alternatively, we consider a second model that exploits the temporal trajectories of the modulation tensor $\mathcal{R}$. In this case, instead of integrating $\mathcal{R}$ over the audio segment, the temporal trajectories of $\mathcal{R}$ across multiple time frames over the duration of the audio segment are modeled. Here, we contrast two approaches to modeling these temporal dynamics. First, we explore the commonly-used derivative features that concatenate the base features with their respective first (Δ) and second derivative ($\Delta\Delta$) components [82]. In this case, the mean, Δ and $\Delta\Delta$ features are computed from each audio segment $\mathcal{R}$ and concatenated to generate 1260 dimensional feature vector for building GMM models. The statistical models based on this feature representation exploit some degree of information contained in the temporal dynamics of modulation features.

Alternatively, we explicitly model the temporal trajectories of $\mathcal{R}$ using a Hidden Markov Model (HMM) framework [20, 83]. Each audio segment of duration 1 second is divided into fixed number of frames of duration t_δ (t_δ= 16 ms) to obtain a time series. Then, HMM models parameterized by π_s,

$P(s_{t+1}|s_t)$ and $P(y_t|s_t)$ are built where s_t denotes hidden states and y_t denotes the actual observation emitted by hidden state at time instant t. π_s denotes the prior distribution of states, $P(s_t|s_{t-1})$ denotes the transition probability matrix and $P(y_t|s_t)$ is the distribution of observations emitted by hidden states mainly modeled as a Gaussian. The hidden states used in our HMM set up represent which of the frequency channels are active at a particular time instant and the transition of one state to another state corresponds to how the activity of one frequency channel changes with respect to other channels over time. The parameters π_s, $P(s_t|s_{t-1})$ and $P(y_t|s_t)$ of the HMM are learned using the Baum-Welch (BW) algorithm as described in [84].

2.3.3 Fusion of GMM-HMM models

We also investigate a hybrid model that combines both mean modulation statistics obtained from the GMM model with the temporal trajectories tracked by the HMM model. Here, the underlying assumption is that both models provide complimentary information that gives an even better representation of intricate changes and dynamics in a sound class, that each model by itself would fail to capture. The proposed hybrid GMM-HMM system operates by combining the GMM and HMM models for each sound class $c_1, c_2, \ldots, c_k$ using a logistic regression [85]:

$$C = \underset{c_1,c_2,\ldots,c_k}{\arg\max}\, w_{GMM} L_{GMM} + w_{HMM} L_{HMM} \tag{2.3}$$

where C is the class to which the test sample gets assigned, L_{HMM} and L_{GMM} are the respective normalized likelihood scores obtained using HMM and

GMM models against a test sample. The logistic weights are trained using a subset development set from the database.

2.4 Experimental Setup and Results

2.4.1 Data

The scene recognition experiments are performed on entire dataset from the BBC Sound Effects Library [76]. The database has total of 2400 recordings, amounting to 68 hours of data. The recordings are organized into 17 classes, for example Ambience, Animals, Transportation and Musical etc. We resample each of the recordings in the database to 16 KHz and preprocess them through a pre-emphasis filter with coefficients $[1 \ -0.97]$ in order to boost high frequencies. 80 % of recordings are randomly selected from the database and used as training set. The remaining 20 % are divided into test and development sets. This latter set is used to train the logistic regression model for the hybrid system. We run a 7-fold cross validation on the entire dataset and report mean accuracy and standard deviation across runs.

2.4.2 Baseline Setup

The proposed system is contrasted against a baseline setup using MFCC features along with their derivative Δ and $\Delta\Delta$ components. Such setup is close to that used in [22]. We compute 13 MFCC features for every frame size of 25 ms with 10 ms overlap. The average statistics, first and second order delta components of MFCC features are computed across these time frames over a duration of 1 second and concatenated to form a 39 dimensional vector. These

vectors are then used to build GMM models for each sound class.

2.4.3 Results and Analysis

Table 1 summarizes the scene classification accuracy using our proposed hybrid system as well as other setups. The results compare the modulation features against the standard MFCC features along with their derivatives (Δ, $\Delta\Delta$). The performance of individual GMM and HMM classifiers using modulation features and their delta components are also reported to assess their respective accuracy values.

Features	Classification Accuracy (%)
GMM based MFCC + Δ + $\Delta\Delta$	49.8 ± 9.5
GMM based modulation features	64.6 ± 5.8
GMM based modulation features + Δ + $\Delta\Delta$	66.8 ± 5.1
HMM based modulation features	65.3 ± 6.4
GMM-HMM based modulation features	$\mathbf{76.57 \pm 4.3}$
GMM-HMM based modulation features + Δ + $\Delta\Delta$	$\mathbf{79.1 \pm 4.1}$

Table 2.1: Results obtained using different features and modeling approaches on Scene Classification Task. $\pm$ indicates the standard deviation across folds.

A number of interesting observations are worth noting. Firstly, the modulation-based features provide a clear advantage over MFCC features in capturing scene characteristics; even with use of derivative components. Secondly, the use of derivative components with modulation features further improve the accuracy of classification suggesting that temporal dynamics captured in

the rate modulation analysis do not sufficiently represent broader temporal changes in the signal that can be better modeled using derivative cues. Thirdly, the HMM system is slightly worse than the GMM system with the derivative features indicating that the mean statistics captured by the modulation features and their dynamics are likely capturing key aspects of each scene that are not well modeled by the HMM system. Consequently, the hybrid system does provide noticeable improvement further corroborating the observation that representing the average distribution of the features with sufficient statistics complements the temporal trajectories in best modeling heterogeneity in sound classes in the BBC dataset.

In order to gain a greater insight into the contribution of each of the GMM and HMM models, we examine the performance of these classifiers for each class of scenes using a detection measure of d' [86]. d' is a very popular measure of sensitivity in signal detection theory (SDT) mainly measured in terms of Hit rate (H) corresponding to number of times the model correctly classifies the test signal and False Alarm rate (FA) equal to number of times the model assigns the test signal to wrong class. d' is calculated as : $d' = z(FA) - z(H)$, where z(FA) and z(H) indicate z scores of false alarm and hit rate respectively. Higher value of d' for a class indicates that the model has a high probability of correctly classifying the test signal, hence a model's classification performance can be well represented in terms of its d' value. Figure 3.2 shows the d' values broken down by class. It is worth noting that most scenes do exhibit an improved accuracy using the hybrid HMM-GMM system. However, such improvement is not noted across *all* classes. One possible reason for decreased performance of the hybrid system for some sound classes could be due to

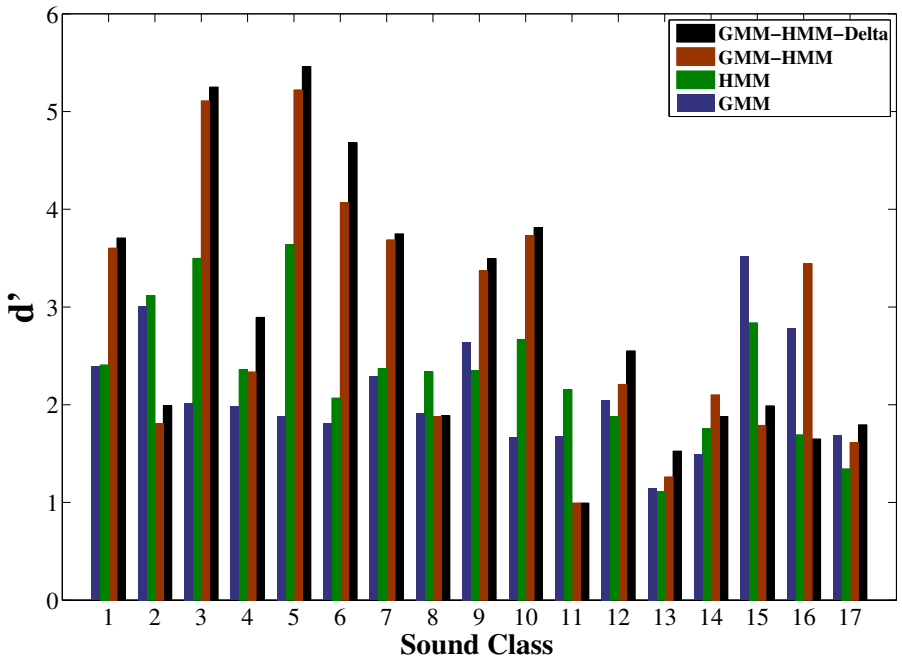

Figure 2.2: *Classification accuracy of various classifiers in terms of d'.* Sound classes used in the classification task : 1. Ambience 2. Animals 3. Emergency 4. Fire 5. Foley 6. Household 7. Humans 8. Impacts 9. Industry and Machines 10. Musical 11. Science Fiction 12. Sports 13. Technology 14. Transportation 15. Warfare 16. Water 17. Weather

the greater heterogeneity in those subsets which undermines the score fusion using a simple logistic model. Another interesting observation is the performance of the GMM vs HMM systems across different classes. For instance, the HMM classifier clearly outperforms the GMM for a class like 'Musical' which includes different tones with varying degree of spectral and temporal modulations. The temporal characteristics of melodies in this class appear to be best represented using the HMM model; in contrast with a class such as 'Water' for instance.

Generally, musical signals have a rich temporal structure and exhibit high

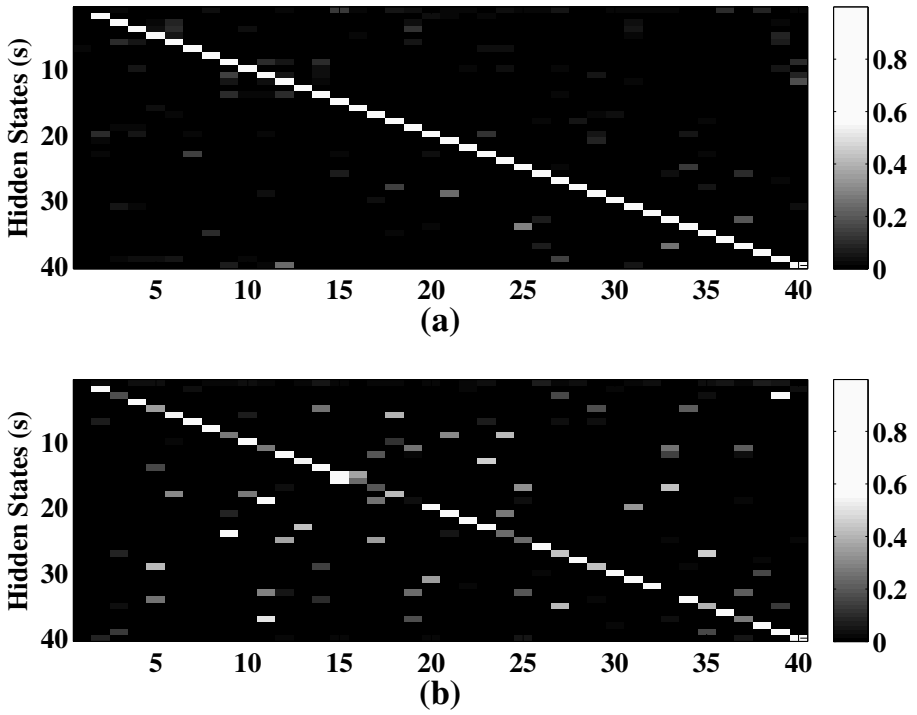

Figure 2.3: Probability transition matrix ($P(s_t|s_{t-1})$) for class (a) Musical and (b) Water

degree of temporal regularity [87]. HMM models capture these 'hidden' regularities in a much effective manner than GMM by means of its probability transition matrix. Figure 2.3 depicts the HMM model's probability transition matrix for 40 states corresponding to classes 'Musical' and 'Water'. In 'Musical', there is a very strong activity across the diagonal elements of the transition matrix which means the frequency channels tend to remain in their own state across multiple time frames corresponding to their strong temporal regularity. In case of 'Water' , the non zero probabilities in non-diagonal elements of the matrix show that the frequency channels tend to make rapid transitions across each other which affects the temporal structure of the scene.

However, because of complimentarity of the information present in temporal structure and average statistics of the scenes, the combination of GMM and HMM models via model fusion gives a tremendous boost in classification accuracy of both the classes as shown in Figure 3.2.

2.5 Discussion

In this chapter, we examine the role of temporal dynamics of modulation features in capturing intricate details in auditory scenes that extend beyond average statistics of the scene and track the heterogeneous dynamics commonly encountered in these scenes. Specifically, we propose that temporal trajectories of local spectral and temporal profiles do provide complimentary information in addition to their mean statistics. A fusion system based on both representations provides a better model of each sound class relative to the individual models. Such hybrid modeling is crucial in case of complex and unconstrained recordings such as the BBC sound effects data. It is common in such datasets that audio samples representing a similar nominal class but different scenarios are grouped under the same label. Modeling these disparate settings requires not only a representation of the characteristics of the sound sources in the scene, but aspects of their temporal dynamics as well. The proposed model based on a hybrid GMM-HMM model along with derivative components provides noticeable improvement over a MFCC-GMM system (by about 30%) as well as individual GMM or HMM systems (by an average of 14%).

One possible limitation of this framework is the imposition of linearity constraint upon spectro-temporal modulation features. Indeed, recent physiological evidence indicates that the representation becomes increasingly complex and nonlinear across the higher stages in auditory pathway [88] To address this issue, we propose a hierarchical framework in the next chapter which learns the spectro-temporal bases from natural soundscapes in an unsupervised and non-linear fashion and explore how such a representation drives scene segregation processes across wide range of auditory scenes.

Chapter 3

A Hierarchical framework for auditory scene segregation

3.1 Introduction

We live in busy environments and our sensory system is continuously flooded by complex information from our surroundings which need to be analyzed in a way that helps interpret the world around us. This process, referred to as scene analysis, is common across all sensory modalities including vision, audition and olfaction [89] and refers to the ability of humans, animals and machines alike to parse the mixture of cues impinging on our senses, organize them into meaningful groups and map them onto relevant foreground and background objects. Gestalt principles are a set of rules that guide this process of scene analysis across sensory systems [90]; and offers guidelines at the core of many theoretical accounts of perceptual organization of scenes common across modalities [91, 92].

In most accounts, the process of scene analysis can be conceptualized as a two-stage process: segregation (or analysis) and grouping (or fusion) [93].

In the first stage, the sensory mixture is decomposed into feature elements, believed to be the building blocks of the scene. The features reflect the physical nature of the sources in the scene, the state and structure of the environment itself, as well as perceptual mappings of these attributes as viewed by the sensory system. These features vary in complexity along a continuum from basic (e.g. edges or frequency components) to more complex features (e.g. shapes or timbral profiles). The ubiquitous nature of these profiles conceals the multiplexed structures that give rise to this analysis. In many models, this stage is represented either in simple feature axes or shaped via complex dimensionality mappings via kernels [94]. Complementing this process is then a fusion stage that integrates the state and behavior of these building blocks to form perceptually cohesive structures forming *objects* or *streams* [90, 95]. These grouping mechanisms reflect the local and global distribution and dynamics of the features and offer 'rules' by which events or sub-events likely arise from the same underlying source or common perceptual object (these two need not always be related in a one-to-one mapping) [1]. In many models, these grouping cues are often leveraged in the context of back-end classifiers that are tuned to capture pattern relationships in specific object classes (e.g. speech, music, faces, cars, etc) [43, 44, 96, 97]. In doing so, these models effectively capture the inter-dependencies between object attributes and learn their mapping onto an integrated representational space [11, 98, 99]. Effectively, success in tackling scene organization depends on two key components [100]: 1) obtaining a rich and robust feature representation that can capture object specific details present in the scene; 2) group the feature elements such that their spatial and temporal associations match the dynamics of objects within

the scene.

Vision models have been very successful in mining these two aspects of scene analysis. Intricate hierarchical systems have leveraged inherent structure in static and dynamic images to extract increasingly elaborate features from a scene that are then used to segment it, interpret its objects or track them over time [101]. Data-driven approaches have shown that high dimensional feature spaces are very effective in extracting meaningful semantics from arbitrary natural images [5, 102, 103]; while hand-engineered features like scale-invariant feature transform (SIFT) [104], histogram of oriented gradients (HOG) [105], and Bag-of-visual-word descriptor [106] among others have also enjoyed a great deal of success in computer vision problems like image classification and object detection. Recent advances in deep layered architectures have resulted in a flurry of rich representational spaces showing selectivity to contours, corners, angles and surface boundaries in images [107–110]. The deep nature of these architectures has also led to a natural evolution from low-level features to more complex, higher-level embeddings that capture scene semantics or syntax [8, 111].

In audition, computational approaches to tackle auditory scene organization have mostly taken advantage of physiological and perceptual underpinnings of sound processing. A large body of work has built on our knowledge of the auditory pathway, particularly the peripheral system to build sophisticated analysis models of an auditory scene. These systems extract relevant cues from a scene, such as its spectral content, spatial structure as well as temporal dynamics; hence allowing sound events with uncorrelated acoustic behavior to occupy different subspaces in the analysis stage [19, 20, 112–115]. These

models are quite effective in replicating perceptual results of stream segregation especially using simple tone and noise stimuli [33, 35–38, 116]. Some models also extend beyond early acoustic features to examine feature binding mechanisms that can be used as an effective strategy in segregating wide range of stimuli ranging from tone sequences to spectro-temporally complex stimuli like speech and music [18, 39, 40]. In most approaches however, the models are built around hand-crafted feature representations, hence limiting their scope to specific mappings of acoustic space. With the emergence of deep belief architectures, recent efforts have started borrowing concepts from vision literature in terms of learning rich feature space from natural soundscape in a data driven fashion, and subsequently using these spaces in domains like music genre classification, phoneme classification and speaker identification [66, 117–120]. Still, these models have mostly been tailored to specific applications via supervised learning techniques; hence limiting their applicability to the problem at hand.

The current study aims to leverage the role of Gestalt principles in the general problem of auditory scene analysis; hence drawing on principles of scene organization that are common across many sound environments. It leverages recent advances in deep learning to examine what kind of cues can one infer from natural sounds; how well do they reflect the known analysis and grouping cues of auditory streams; and how effective are these cues in explaining perceptual organization of auditory scenes with varying degrees of complexity. We propose a multi-layered deep belief neural network designed to analyze the incoming acoustic signal with a multitude of granularities. Besides the basic layout and choice of analysis window sizes, the network is trained in

an unsupervised fashion to explore which cues naturally emerge from a rich set of sounds including speech, nature, etc. The architecture is specifically laid-out in two stages mimicking a segregation process where we expect local and global acoustic attributes to emerge; hence segregating the incoming scene along disparate feature axes. The short-term analysis underlines the tiling of the spectro-temporal space; hence inferring local cues referred to as *simultaneous* grouping cues [121–123]. The longer-range analysis extends the segregation stage to examine temporal dependencies across acoustic attributes over different time scales; hence exploring emergence of *sequential* grouping cues [124–128]. Finally, a fusion stage binds the cues together based on how strongly they correlate with each other across multiple time scales. This integration is achieved using *hebbian* learning framework which reinforces activity across coherent channels and suppresses activity across incoherent ones [129–131].

The overall system is tested with a wide range of stimuli where we can quantify the role of each and every component in the network in driving stream segregation processes. We also contrast the system performance with a set of control experiments where different components of the system are deliberately switched on/off in order to examine their impact in the organization of different acoustic scenes. The paper first presents an in-depth description of the proposed architecture, followed by an analysis of the emergent properties of the trained network and their potential neural correlates in the auditory pathway in section 3.2. The experimental results shown in section 3.3 outline how the network replicates human psychoacoustic behavior in stream segregation and speech intelligibility paradigms. Finally, we present control experiments

that dissect the network architecture and examine the contribution of each and every component. In section 3.4, we discuss the implications of this network in shedding light on ties between observed perceptual performance in various complex auditory scenes and the neural underpinnings of this behavior as implemented in networks of neurons along the auditory pathway.

3.2 Proposed Architecture

3.2.1 A model for sound organization

The proposed model develops a hierarchical network of auditory processing that echoes the infrastructure of early- and mid-audition along the mammalian auditory pathway [132]. The model maps incoming sound signals onto an increasingly higher dimensional space that evolves from shorter, localized time constants to longer scales; conceptually emulating the variations of integration windows along the auditory hierarchy from the periphery up to the midbrain then auditory cortex. Other than these design elements, the model tuning is learned in a data-driven fashion. Being trained on datasets of natural sounds, each component of the model learns in a generative fashion to represent natural sounds from its own vantage point following principles of deep belief networks, as detailed next. Figure 3.1A depicts a schematic of the overall model. It takes as input the acoustic waveform of an auditory scene $u(t)$, which is then mapped onto a time-frequency representation $S(t, f)$ mimicking peripheral processing in the auditory system. This first transformation analyzes the acoustic signal $u(t)$ using a bank of logarithmically-spaced

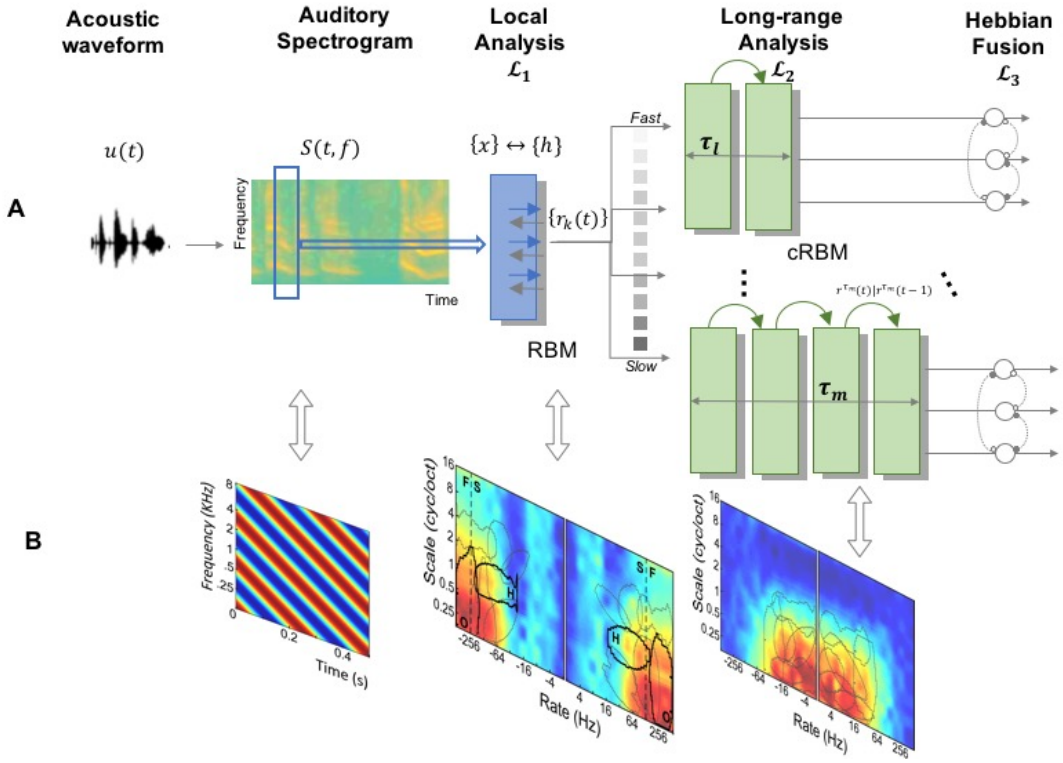

Figure 3.1: Schematic of the proposed model (A) An acoustic signal undergoes a series of transformations starting with a mapping to a time-frequency spectrogram, followed by two-layers of deep neural networks, then a fusion stage. (B) Noise ripples are used to analyze the spectro-temporal tuning of the model at different stages. The modulation transfer function for layers $\mathcal{L}_1$ and $\mathcal{L}_2$ are shown in the rate-scale domain. The frequency axis f is collapsed for display purposes. Overlaid on each transfer function is a contour plot of agglomerative clusters in spectro-temporal modulation space.

asymmetric constant-Q spectral filters. The filter-bank comprises of 128 asymmetric filters equally-spaced on a logarithmic axis over 5.3 octaves spanning the range 180 Hz to 8000 Hz. Next, the signal undergoes spectral sharpening via first order derivative along the frequency axis followed by half wave rectification and short term integration with $u(t,\tau) = e^{-t/\tau}u(t)$ where $\tau = 10$ ms. This filterbank analysis results in a time-frequency auditory spectrogram represented by $S(t,f)$. This analysis replicates the cochlear model of Yang *et al.* [114].

The spectrogram $S(t,f)$ is then sampled over 3 consecutive frames which are grouped together in a process of *shingling* [133]. This step allows the network to perform short term feature analysis over a temporal context of 30 *ms*. This stage (called $\mathcal{L}_1$) is structured as a two-layer sparse Restricted Boltzmann Machine (sparse RBM) with a fully connected visible and hidden layer [111]. The visible layer units x_k are real-valued and characterized by a Gaussian distribution fitted over the input spectrogram $S(t,f)$; while the hidden units h_k are sampled from a Bernoulli distribution such that $h_k \in (0,1)$ for $k = 1,2,\ldots,N$, where N is the number of nodes in each layer. In the current implementation, we set $N = 400$. The network is parameterized by $\theta = W, b_v, b_h$ where W represents the interconnected weights between $\{x\}$ and $\{h\}$, and b_v (b_h respectively) represents the visible (hidden, respectively) bias. The network is trained using the Contrastive Divergence (CD) algorithm with the objective to minimize the reconstruction error between x and $\hat{x} = hW + b_v$ [134]. After training, the connection weights W_k are transformed into a 2D representation capturing localized spectro-temporal filters $\mathcal{F}_k(t,f)$, akin to spectro-temporal receptive fields recorded in the mammalian auditory system

[135]. These learned filters are then applied in a convolutional fashion over the incoming spectrogram $S(t, f)$ to derive the response of each filter. Next, this response undergoes a neural adaptation stage that allows to strengthen the contrast between foreground and background units. This adaptation imposes an exponential decay over each filter response hence suppressing units with weak activation. The final activation of $\mathcal{L}_1$ nodes is then an array of responses $r_k(t)$ which are then processed through the next layer in the hierarchy. $\mathcal{L}_1$ filter responses $\{r_k(t)\}$ form the input to the next layer in the hierarchy ($\mathcal{L}_2$); which is structured to highlight the contextual dependencies in the signal features. This temporal context is captured using conditional RBMs (cRBM), which are extended versions of RBMs designed to model temporal dependencies [136]. Similar to a RBM, a cRBM comprises a visible layer with units x_k, sampled from a Gaussian distribution fitted over the input, and a hidden layer with h_k units sampled from a Bernoulli distribution. Unlike a RBM, a cRBM acts as a dynamical system operating over an entire input history τ taking as input occurrences at times $\{t, t-1, \ldots, t-\tau\}$ in order to capture dynamics in the input space over a context τ. The proposed model multiplexes the outputs $r_k(t)$ from $\mathcal{L}_1$ over multiple histories; hence transforming $\mathcal{L}_1$ activations into multi-rate inputs $r_k^{\tau_1}(t), r_k^{\tau_2}(t), \ldots, r_k^{\tau_K}(t)$ over a range of temporal resolutions $\tau_k \sim$ (30 - 600 ms). $\mathcal{L}_2$ is parameterized by $\theta = U, z_v, z_h, A_\tau, B_\tau$ where U represents the interconnected weights between visible units x and hidden units h, z_v represents the visible bias, z_h represents the hidden bias, A_τ and B_τ characterize the autoregressive weighs between past inputs, the current input and current hidden unit respectively. Since the multi-rate versions of $r_k^\tau(t)$ are instantiations of different acoustic cues across multiple time resolutions τ,

the weights U capture the interactions across these cues over different temporal resolutions whereas the autoregressive weights A_τ and B_τ capture the effect of long term temporal dependencies over such interactions. Just like the localized layer $\mathcal{L}_1$, the contextual layer $\mathcal{L}_2$ is trained in a generative fashion using contrastive divergence (CD) in order to best capture the dynamics in natural sounds using a large dataset of diverse realistic sounds spanning speech, music and natural sounds. Once trained, the model parameters θ are then applied over incoming $\mathcal{L}_1$ filter response in a linear fashion, yielding a multi-resolution output which is then passed over to the next stage in the hierarchy.

The next stage of the hierarchy performs a binding operation across outputs of the $\mathcal{L}_2$ layer, giving rise to perceptually-coherent object representations. It explores co-activations across all the channels within a given context τ_k and binds together the units that exhibit strong temporal coherence. The *'temporal coherence'* theory posits that emergence of perceptual representations of auditory objects depends upon *strong* coherence across cues emanating from same object and *weaker* co-activation across cues from competent objects [137, 138]. This coherence is not an instantaneous correlation but one that is accumulated over longer time scales, commensurate with the contextual windows explored in the $\mathcal{L}_2$ layer. We implement this concept in a biologically-plausible fashion via mechanisms of Hebbian learning, which suggests that when two neurons fire together, their synaptic connection gets stronger [139]. Effectively, Hebbian interactions operate by reinforcing activity across coherent channels, hence grouping them into putative objects and inhibiting activity across incoherent channels [140]. We implement a synaptic interaction across output channels

from layer $\mathcal{L}_2$ by introducing a synaptic weight matrix V. If two units i and j are co-activated at a given time t, their corresponding synaptic V_{ij} is reinforced over time. If the correlation between their activity is weak, the corresponding synaptic weight V_{ij} reduces as well. These synaptic weights are applied to the output of each channel in a dynamic fashion, hence modulating the activity across an entire ensemble of neurons within each context in layer $\mathcal{L}_2$. The net effect gives emergence to perceptual coherent groups that represent auditory objects in a scene.

3.2.2 Model characterization

In order to examine the emergent sensitivity of learned layers in the network, we derive the tuning characteristics of individual nodes or neurons and explore their filtering properties in the modulation domain [141, 142]. Modulation tuning reflects the signal cues that best drive individual nodes in the model both in terms of temporal variations and dynamics (i.e. temporal modulations) as well as spectral span and bandwidth (i.e. spectral modulations). This approach follows common empirical techniques used in electrophyisology and psychophysics to probe the tuning of the system to specific acoustic cues. It is frequently used in characterizations of spectro-temporal receptive fields (STRFs) which offer 2-dimensional profiles of filtering characteristics of a neuron [135].

First, we employ a classic transfer function characterization method using probe stimuli in order to derive the tuning of both $\mathcal{L}_1$ and $\mathcal{L}_2$ layers of the network [143–145]. We present spectro-temporally-modulated noise signals

(called ripples) as input to the model with varying spectro-temporal modulation parameters and we characterize the fidelity of the ripple encoding at various stages of the network as we vary the ripple modulation parameters [146]. Each ripple is constructed as a broadband noise signal whose envelope is modulated both in time and frequency according to the equation:

$$S_{rp}(t, f) = L(1 + \Delta A sin(2\pi(\omega t + \Omega f) + \phi)) \tag{3.1}$$

where L denotes the overall level of the stimulus, t is time, and f is frequency on a logarithmic scale, ω is the temporal modulation or ripple velocity (in Hz), Ω is the spectral modulation or ripple density (in cyc/oct), and ϕ is phase of the ripple (Figure 3.1B-left).

We vary the ripple parameters over the range $\omega \in [-512, 512]$Hz and $\Omega \in [0.25, 16]$ cycles/oct, and compute the modulation transfer function (MTF) from the response of layers $\mathcal{L}_1$ and $\mathcal{L}_2$. Figure 3.1B-middle,right depicts the MTF derived from both $\mathcal{L}_1$ and $\mathcal{L}_2$. The functions highlight that layer $\mathcal{L}_1$ is a faster layer tuned to rapid temporal dynamics, with a clear concentration of energy in $|\omega| \in [64, 512]$; as well as spectral dynamics $\Omega \in [0.25, 4]$ cycles/oct. In contrast, the contextual layer $\mathcal{L}_2$ is mostly tuned to slower dynamics $< 30Hz$ with tighter spectral selectivity mostly concentrated below 1 cycles/oct. This outcome is very reminiscent of similar transfer functions obtained from neurophysiological data showing contrasting tuning characterizations in the midbrain, auditory thalamus and auditory cortex [88, 147, 148], whereby selectivity of individual neurons along the mammalian auditory hierarchy evolves from faster to slower temporal dynamics and from more

refined to broader spectral spans along frequency. The tuning characteristics obtained from the model exhibits a similar behavior reflecting the progression from localized analyses in time and frequency in neurons in $\mathcal{L}_1$ to broader, more context-sensitive selectivity in neurons in $\mathcal{L}_2$.

We further examine the selectivity of *individual* neurons by exploring emergent tuning characteristics common across nodes in the network. We employ an agglomerative clustering algorithm [149] over individual response functions represented as $\mathcal{F}_k(t, f)$ in $\mathcal{L}_1$. This approach clusters nodes exhibiting similar tuning characteristics into common groups hence highlighting selectivity of neuron subgroups from which we can infer a link with underlying acoustic cues being processed. Figure 3.1B shows contour plots from the resulting clusters overlaid on the spectro-temporal modulation space. Neurons in $\mathcal{L}_1$ appear to naturally group around specific modulation regions; giving rise to a wide range of selectivities. Of note, neurons clustered in a group labeled **O** appear to be more sensitive to fast transients or 'Onsets'; while spectrally structured neurons labeled **H** are centered around spectral modulations $\in$ [1-2] cyc/oct corresponding to harmonic peaks present in natural sounds. Other groups also emerge with special selectivity to spectral or temporal features as well as oriented spectro-temporally selective clusters, likely tuned to detect frequency-modulated sweeps in the signal. In addition, there is also a natural grouping of neurons into two broad classes; a fast group (**F**) consisting of neurons with temporal modulations $> 100Hz$ and a slow group (**S**) comprised of neurons responding to modulations $< 100Hz$. We employ similar clustering algorithm over $\mathcal{L}_2$ filters and see that model neurons in the second layer naturally cluster around slower spectro-temporal modulations as shown in Figure

3.1B. This analysis further demonstrates the context-sensitive selectivity in $\mathcal{L}_2$ neurons and justifies the evolution of tuning characteristics in the network from faster, more refined to slower, much broader in the modulation space.

3.3 Experimental Results

3.3.1 Primary Results in streaming and speech intelligibility paradigm

We test the model's behavior with a variety of acoustic scenes ranging from classic streaming paradigms using simple tones to experiments using speech signals. Crucially, all experiments are tested on the *same model* (after all layer have been trained), without any adjustment to model parameters. The stimuli parameters are carefully chosen to closely replicate human perceptual results hence allowing a direct comparison between the model and human perception. All stream segregation results are shown in Figure 3.2 organized in 3 columns: the stimulus on the left, a replica of human perception of the same stimulus reproduced from the corresponding publication in the center, and the model performance on the right.

Simple Tones

The first experiment employs the classic two-tone paradigm with sequences of high and low notes, commonly used in streaming experiments [1, 150, 151]. The sequences are produced by presenting two tones of different frequencies,

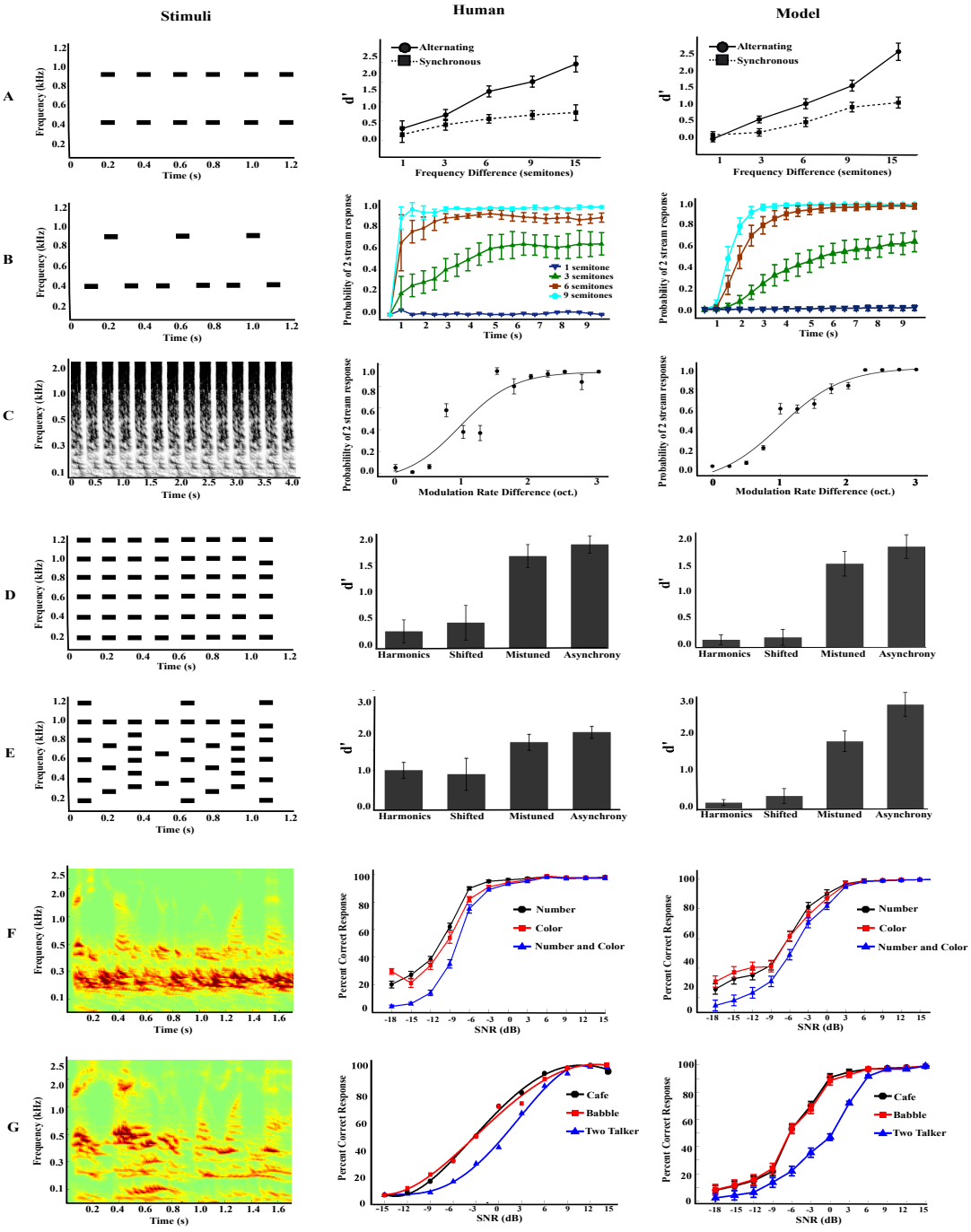

Figure 3.2: Primary results of stream segregation using proposed model. Leftmost panel shows the stimuli sequence used for each experiment. Middle panel shows the human listening performance whereas rightmost panel shows the model performance

A and B, repeatedly and in alternation ($ABAB-$). When the frequency separation ΔF between the A and B tones is relatively small (<10%), listeners perceive the sequence as grouped or fused and report hearing one stream. As the frequency separation ΔF increases, listeners hear two separate streams consisting on only the low notes ($A - A-$) or only the high notes ($-B - B$-). In contrast, when the two notes A and B are presented synchronously (Figure 3.2A-left), listeners tend to hear the sequence as grouped regardless of the frequency separation ΔF, in a process reminiscent of temporal coherence which fuses together channels that are co-activated together [137, 138]. Figure 3.2A-middle replicates results from a study by Micheyl *et al.* [152]. The study shows that an alternating tone sequence is perceived as a single stream when the frequency separation ΔF is small and is segregated into two streams when ΔF is large. When the two tones are presented synchronously, they are always perceived as grouped regardless of frequency separation. The fused percept is objectively measured using d' [153, 154]; where listeners are asked to detect a change in one of the tones presented in the final burst. Figure 3.2A-right shows that the model replicates the same behavior using the same tone sequences presented in alternation or synchrony. As the frequency separation ΔF increases between the A and B tones, the model is more likely to perceive them as segregated in the alternating condition but tends to fuse them in the synchronous condition.

The two-tone paradigm is also often used to probe the phenomenon of buildup of streaming [1, 155]. The buildup highlights that streaming is a dynamic process, whereby the segregation of the two notes into separate streams is not instantaneous; but builds-up over time taking up to several seconds to

emerge. In a study by Micheyl et al. [156], buildup was assessed using a variation of the two-tone paradigm using tone triplets ($ABA - ABA$), as shown in 3.2B-left. Figure 3.2B-middle replicates results from this study [156] whereby listeners *continuously* report perception of one or two streams for different frequency separations ΔF. The behavioral data shows that when the frequency separation ΔF is large, both A and B tones are perceived as segregated streams relatively quickly. As ΔF decreases, the segregated percept takes longer to emerge lasting over many seconds. Figure 3.2B-right replicates the same behavior using the model and shows that the sequences gradually segregate into separate streams with different time constants. The model faithfully replicates human performance; demonstrating a faster buildup at large ΔF, slower buildup at intermediate ΔF, and no buildup at very small ΔF.

Complex Tones

Next, we explore stream segregation using complex tones. These complexes highlight the wide range of acoustic cues that aid in the segregation of auditory scenes; including frequency separation (as shown earlier), as well as amplitude modulations (AM), harmonicity, temporal synchrony, etc. [91, 125, 157, 158]. In this simulation, we focus on the role of modulation cues in stream segregation by replicating a classic study by Grimault et al.[159] where alternating noise bursts with different AM rates are presented (Figure 3.2C-left). As the difference in modulation rate ΔAM increases, noise bursts tend to segregate into two streams with distinct AM rates. Once the rate difference ΔAM reaches about 2 octaves, the modulated noises fully segregate into two distinct

streams. Figure 3.2C-middle shows human perception of segregated streams as a function of ΔAM replicating the results from the study by [159]; while Figure 3.2C-right shows the performance of the model on the same stimuli. As shown in the Figure, the model closely replicates human perception as reflected by increase in probability of stream segregation, hence indicating that the emergent tuning of nodes in the model explicitly encodes information about amplitude modulation; hence allowing the model to leverage this information to facilitate segregation of noise sequences into corresponding streams.

Next, we examine the role of harmonicity and temporal synchrony as putative grouping cues. Both these cues are believed to exert strong grouping acting as a bond that fuses sound elements together as shown in study by Micheyl et al. [160]. In this work, a target tone at frequency 1000 Hz is masked by background tones that are either harmonically related or in temporal synchrony with the the target tone. The study examines two kinds of stimuli: 'MBS' -multiple burst same- stimuli (Figure 3.2D-left) have the same burst of tones are presented every time; 'MBD' -multiple burst different- stimuli (Figure 3.2E-left) vary the harmonicity relationship between target and background tones at every burst based on different fundamental frequencies. Figures [3.2D,3.2E]-middle replicate the results from the study by Micheyl et al. [160] in which listeners detect a change in the final burst of the target tone. The study shows that when target and background tones are either harmonically related or in temporal synchrony with each other, d' is low indicating a strong background-target fusion. Listeners' ability to segregate the target improves when either harmonicity or sychrony is perturbed. Figures [3.2D,3.2E]-right

show the model performance on the same MBS and MBD stimuli respectively. When target and background tones are harmonically-related or in synchrony, the model favors fusion and results in a small d'. In contrast, perturbing harmonicity by shifting the harmonics, the model favors a segregated interpretation resulting in increased d'. Similarly, when target and background tones are asynchronous, there is a significant increase in d', again suggesting strong segregation.

Speech Intelligibility

Next, we examine the model's behavior using complex sounds such as speech in presence of competing noise. In all experiments, a speech utterance is presented to the network either in clean or masked by background noise that includes speech modulated noise, babble noise, cafe noises or an interfering speech utterance. All speech utterances are part of the CRM corpus where each utterance consists of a call sign and a colorâĂŞnumber combination, all embedded in a carrier phrase [161]. A typical sentence would be "Ready baron, go to red four now," where 'baron' is the call sign, and 'red'-'four' is the colorâĂŞnumber combination. Figures [3.2F,3.2G]-left show spectrograms of speech utterances from the CRM corpus mixed with speech modulated noise and an interfering speech utterance respectively.

Figures [3.2F,3.2G]-middle replicate the results from two behavioral studies using the CRM corpus in a dichotic listening paradigm where subjects identified the "number" and "color" mentioned in the target utterance under different noise conditions [162, 163]. The behavioral data yield a measure of speech intelligibility (in word percent correct) as a function of signal to noise

ratio (SNR) with different noise maskers. Figures [3.2F,3.2G]-right depict the model's performance replicating the same paradigm as closely as possible. The model yields a correct identification of speech tokens (numbers, colors, or both) that is closely related to the SNR condition following an S-shaped curve typical of similar measures of speech intelligibility in noise. The model performance plateaus at about 98% correct identification at SNRs above 3dB (Figure 3.2F-right); whereas it degrades quite rapidly from -3 to -9 dB before reaching chance performance at -18 dB. When comparing effects of noise type, both human and model performance is poorer in presence of an interfering utterance, relative to babble and cafe noise conditions.

3.3.2 Model function and malfunction

As outlined earlier, Figure 3.2 outlines how the model is able to faithfully replicate a wide range of perceptual results for stream segregation. Next, we examine the actual contribution of different components of the model to get a better perspective on how each principle modeled in the system contributes to stream segregation. The experimental results shown in the previous section suggest that simultaneous cues (tonotopic organization, AM rate, harmonicity, temporal synchrony, etc), sequential cues and grouping mechanisms play an important role in streaming paradigms. In order to shed light on their individual contributions, we run a series of *control* experiments where we look at malfunctions in the model if components of the system are disrupted individually.

Role of Simultaneous cues

The tuning characteristics of layer $\mathcal{L}_1$ show that model neurons naturally cluster around specific modulation regions, hence, revealing a wide selectivity to different acoustic cues that emerge in natural sounds. Here, we focus on four $\mathcal{L}_1$ neuron clusters with particular selectivity to harmonicity, onsets, fast and slow temporal modulations. We individually 'turn off' each of these clusters from the system and replicate all stream segregation experiments shown earlier. Figure 3.3 shows the model performance as follows: The left-most column shows the model performance when $\mathcal{L}_1$ harmonicity-neurons are turned off, the middle column with $\mathcal{L}_1$ onset neurons turned off, and the rightmost column with fast ($>$ 100 Hz) and slow ($<$ 100 Hz) $\mathcal{L}_1$ units turned off respectively. In these experiments, $\mathcal{L}_2$ is not altered but is retrained based on a modified input (i.e. its input dimensionality is reduced because harmonicity, onset, slow or fast channels are removed).

Switching off harmonicity-$\mathcal{L}_1$ nodes has no effect on the system's performance in a two tone paradigm (Figure 3.3A-left) or sinusoidally amplitude-modulated noise bursts (Figure 3.3B-left). In contrast, the ability to segregate MBS and MBD sequences in case of mistuned harmonics is drastically affected by the absence of harmonicity-tuned nodes in the network (Figure 3.3C,D-left). Similarly, the network's ability to detect speech (i.e. color and number in the CRM corpus) is severely impacted in absence of harmonicity-tuned nodes (Figure 3.3E-left). Taking a closer look at the behavior of the network in detecting numbers, we note a systematic drop in performance across all digits

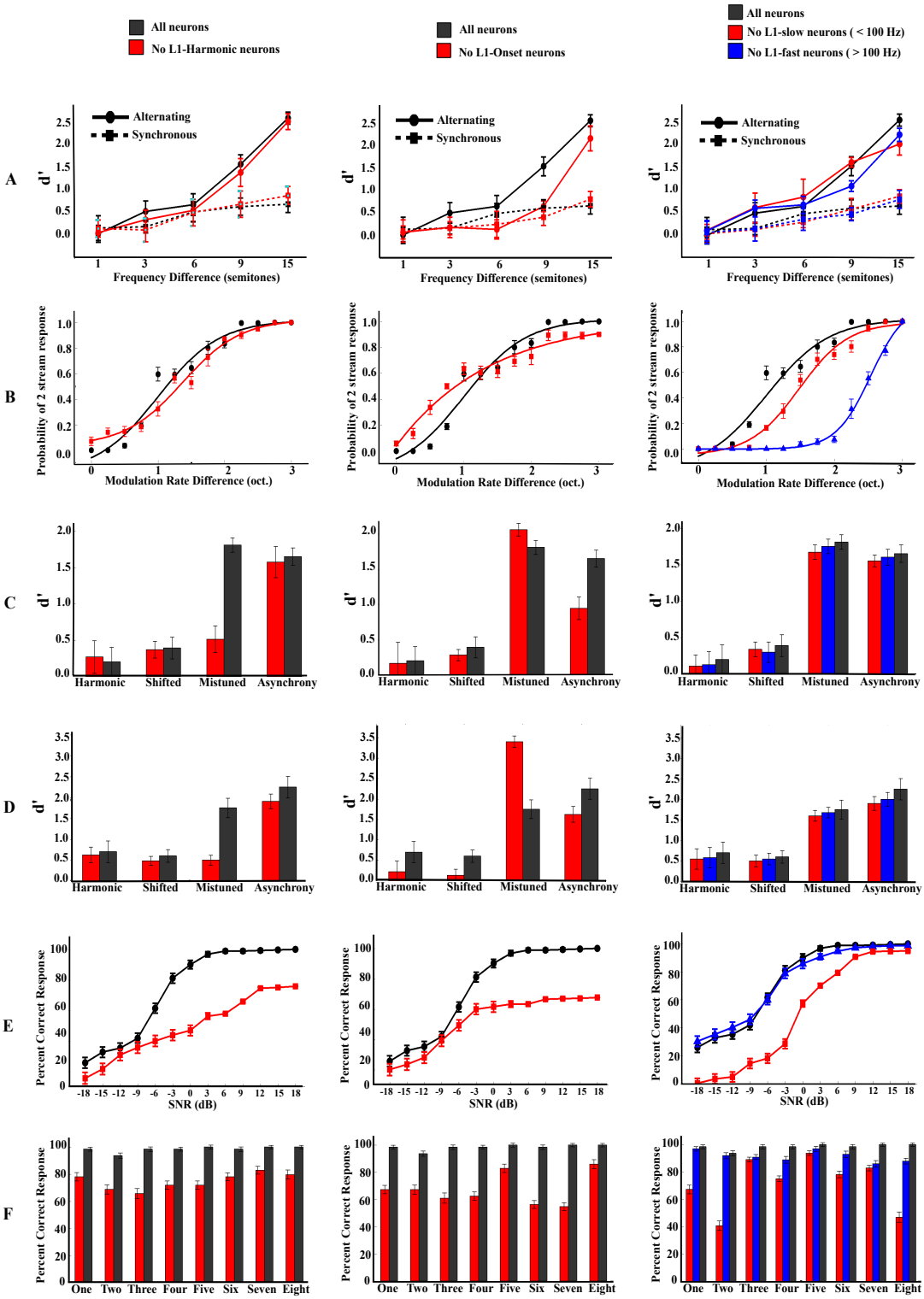

Figure 3.3: Control experiments introducing malfunction in layer $\mathcal{L}_1$

which all contain prominent voiced phonemes (Figure 3.3F-left). A similar systematic drop is also noticed across all color key words in the corpus (data not shown).

Disabling $\mathcal{L}_1$-onset nodes results in its own malfunctions of the model. Streaming two-tone sequences and sinusoidally amplitude-modulated noise bursts is not affected by switching off onset units (Figure 3.3A,B-middle). However, the MBD and MBS stimuli appear to be affected in an interesting way (Figure 3.3C,D-middle) where we note an improvement of segregation in case of mistuned harmonics. The design of these stimuli puts temporal synchrony and harmonicity in conflict. Free of onset-detectors, the model is able to judge segregation mostly driven by harmonicity or lack thereof in the case of mistuning. Conversely, in case of temporal asynchrony, there is a drop in segregation performance in absence of onset-detectors, though the model is able to exploit the harmonic relationship between target and background tones to induce streaming. A comparable drop in speech intelligibility performance is also noted (Figure 3.3E-middle), attesting to the important role of onsets in speech perception. Taking a closer look at the model performance with individual digits (Figure 3.3F-middle), we note severe drops for tokens like “three", “six" and “seven" that contain prominent fricative and plosive unvoiced phonemes. The role of amplitude modulations in stream segregation has a different effect on the model's behavior. We manipulate the selectivity of $\mathcal{L}_1$ neurons to different range of amplitude modulations by testing only-slow or only-fast neurons ($<$ or $>$ 100Hz respectively)). The segregation of two-tone sequences appears to be unaffected by presence or absence of slow or fast units alone, and is likely mostly driven by the tonotopic organization of the nodes in the

network (Figure 3.3A-right). In contrast, streaming of sinusoidally-modulated noise bursts is heavily affected when $\mathcal{L}_1$ units tuned to faster modulations are turned off, though only mild changes are noted when slower-units are turned off (Figure 3.3B-right). Streaming of MBD and MBS sequences appears unaffected by the time-constants of temporal modulations left in the $\mathcal{L}_1$ layer; and we observe no changes to the model behavior (Figure 3.3C,D-right). Interestingly, speech intelligibility is also unaffected when faster $\mathcal{L}_1$ units are turned off (Figure 3.3E-right). In contrast, switching off slower units drastically affects the model's ability to separate speech from noise, especially at low SNRs, strongly corroborating the role of midrange-modulations in speech perception [142].

Role of sequential temporal dynamics

Next, we examine the impact of model parameters responsible for temporal integration on stream segregation over longer time scales. First, we observe the model's behavior if we switch off neural adaptation at the output of $\mathcal{L}_1$ nodes. This mechanism aims to adjust the dynamics of neurons' responses by eliminating nodes with moderate activation over time. Figure 3.4-left contrasts the model's performance with and without this neural adaptation. Figure 3.4A-left shows that neural adaptation is important for segregating alternating two-tone sequences. Adaptation appears to aid the temporal coherence layer in 'shutting down' neurons from competing streams which facilitates segregation. In its absence, both tones in the stimulus continue to compete at the output of the model hence affecting the ability to segregate. Furthermore,

this continued competition appears to slow-down the buildup process (Figure 3.4B-left compared to the original model behavior in Figure 3.2B-right). As noted in the figure, a tone sequence with frequency separation of $\Delta F = 9$ semitones takes many seconds to eventually reach a segregated percept with modified model as compared to 1-2 secs in the original model, owing to the continued competition between the two tones. While the temporal coherence model is able to note the out-phase relationship between the streams, this process is assisted by neural adaptation which supresses activity from competing streams hence speeding up stream segregation in line with observed behavioral responses (Figure 3.2B-middle). A similar behavior is observed in case of sinusoidally amplitude-modulated noise bursts in Figure 3.4C-left. Here again, removing adaptation from the network allows competition across channels to linger longer hence hampering the role of temporal coherence in detecting consistent incoherent activity across competing streams. In the case of MBD and MBS sequences, adaptation appears to have a mild effect with the exception of mistuned harmonics in the case of MBD sequences and temporal asynchrony for MBS sequences (Figure 3.4D,E-left).

We next explore the role of temporal dynamics in cue extraction, particularly the role of slower time-constants which are thought to play a crucial role in sequential integration of acoustic cues as the scene evolves. We probe this role in a control experiment by switching off the $\mathcal{L}_2$ units with strong selectivity to modulation rates (< 50 Hz) and compare this modified network against the full architecture. The results comparing the two models are shown

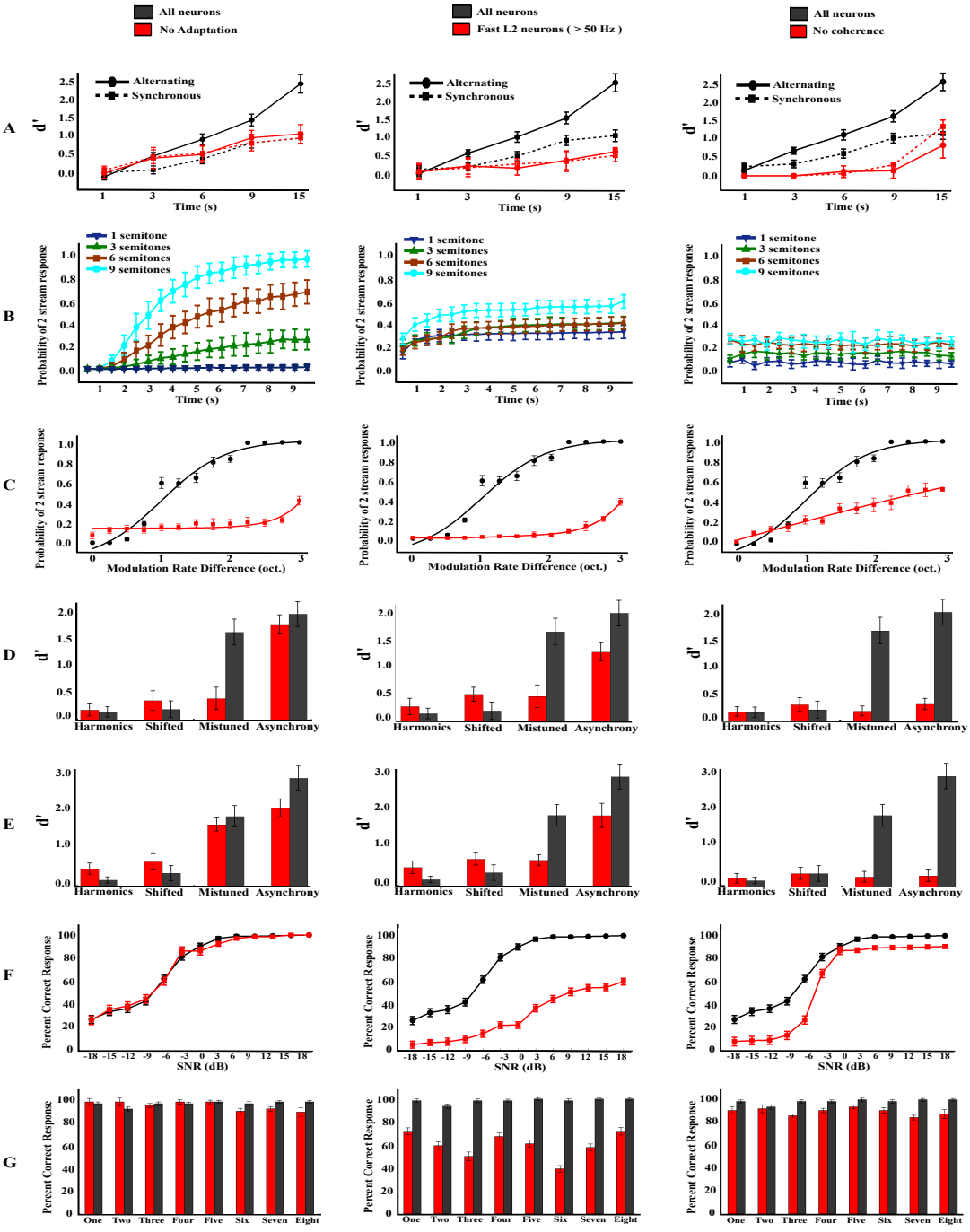

Figure 3.4: Control experiments introducing malfunction in temporal dynamics of the network

in Figure 3.4-middle and reveal wide spread aftereffects across all streaming experiments. In the case of the two-tone paradigm, removing slower neurons from $\mathcal{L}_2$ significantly impairs the network's ability to segregate 2 streams as ΔF increases (Figure 3.4A-middle). Also of note is that the streaming buildup is severely affected and quickly settles on final assessment of segregation between streams regardless of ΔF value likely reflecting the inherent spectral-based separation across the neurons in the network but failing to track how activity across the neural population evolves over time (Figure 3.4B-middle). Segregation of modulated noise bursts is also severely affected (Figure 3.4C-middle). The probability of perceiving 2 streams drops dramatically, indicating a poor integration of neural activity across differentiated neurons. The same effect is observed in the case of MBS and MBD sequences, where the network fails to segregate the target tone from background masker tones even in presence of mistuned harmonic relationships (Figure 3.4D,E-middle). This drop is also noted for both stimuli in the case of asynchrony, even though the drop is not as dramatic, suggesting the network still relied on some degree of temporal alignment across the fast neurons remaining in the $\mathcal{L}_2$ network to judge relationship between tone bursts. Finally, in the case of speech in noise experiments, the network containing 'faster' neurons only is severely impaired across all SNR values (Figure 3.4F-middle). The drop in performance is clear across all digits (Figure 3.4G-middle), as well as across colors (data not shown). The absence of slow $\mathcal{L}_2$ units clearly affects the network's ability to match the slow changes in temporal structure of speech tokens even in presence of simultaneous cues hence failing to facilitate stream segregation.

Finally, the role of temporal fusion across channels is examined by testing the model's performance without the temporal coherence mechanism in layer $\mathcal{L}_3$. Much like earlier control experiments, removing temporal coherence has sweeping effects on the model's ability to perform stream segregation. In the two-tone paradigm, the model treats the synchronous and alternating notes similarly as it fails to judge the phase relationship across spectral channels (Figure 3.4A-right). The buildup of streaming is also completely annihilated regardless of frequency separation across channels strongly suggesting that integration over time and across frequency channels plays an important role in the brain's ability to consolidate information spectrally and temporally while it examines possible configurations or interpretations of the scene (Figure 3.4B-right). This process is very much what the temporal coherence stage achieves and is clearly impaired without coherence. Segregation of modulated noise bursts is also affected although the probability of segregation does increase with increased AM rate difference ΔAM albeit with reduced probability suggesting poorer segregation performance of the modified network (Figure 3.4C-right). In the case of noise complexes in the MBD and MBS paradigm, the network completely fails to achieve any form of segregation (Figure 3.4D,E-right) suggesting that the presence of simultaneous cues (e.g. harmonicity) is not sufficient. Complex noise patterns tend to activate a wide range of channels which require an integration mechanism such as $\mathcal{L}_3$ temporal coherence to interpret based on across-channel consistency and phase relationships. Speech segregation is slightly affected by the disabling of temporal coherence (Figure 3.4F-right) and more noticeably at lower SNR values. Mild reductions in segregation are observed consistently across all digits (Figure 3.4G-right)

and colors (data not shown).

3.4 Discussion

This study presents a biologically-plausible model of stream segregation that leverages the hierarchical and non-linear representation of sound along the auditory pathway. While the model is formulated to focus on local and global cues in everyday sounds, it is structured so that it 'learns' these cues directly from the data. The unsupervised nature of the architecture yields physiologically and perceptually meaningful tunings of model neurons that support the organization of sound into distinct auditory objects. The three key components of the architecture as shown in Figure 3.1 are : (1) A deep belief **RBM layer** that encodes two-dimensional input spectrogram into *localized* specto-temporal basis representation based on short term feature analysis; (2) A dynamic **cRBM layer** that captures the long-term temporal dependencies across spectro-temporal bases characterizing the transformation of sound from fast changing details to slower dynamics. (3) A **temporal coherence layer** that mimics the hebbian process of binding local and global details together to mediate the mapping from feature space to formation of auditory objects.
The layout of the model closely replicates the physiological layout of auditory processing in the brain where an acoustic signal undergoes a series of transformations from the cochlea all the way to auditory cortex (A1), effectively extracting a rich feature representation [15–17, 164–166]. This multitude of transformations evolves in temporal and spectral resolution going from temporally fast, spectrally refined as is typically observed at the level of the

midbrain to markedly slower and spectrally broader and richer in cortical networks [147, 167, 168]. The model 'learns' similar structures as can be seen from the modulation transfer functions for both layers $\mathcal{L}_1$ and $\mathcal{L}_2$ (Figure 3.1B). This feature representation is complimented with fusion mechanisms that give rise to perceptually coherent objects. Temporal coherence has been shown to play an important role in this binding process, operating on the feature space to integrate cross-channel activity, and has been speculated to operate beyond auditory cortex likely in a network engaging the intraparietal sulcus and superior temporal sulcus [169–172]. The proposed model captures the role of temporal coherence in the final layer $\mathcal{L}_3$ via hebbian inetractions.

Role of Simultaneous Layer: Extracting relevant information from incoming acoustic waves is the backbone of any further processing and sound interpretation. The model replicates this feature analysis in a data-driven fashion by employing a diverse dataset of natural sounds including human speech, animal vocalizations and street ambient sounds. Structuring the local layer using an RBM architecture allows the model to learn a rich tiling of spectro-temporal basis functions. The results indicate that these bases capture fine details in the acoustic stimulus, as suggested by the modulation transfer function (Figure 3.1B); showing a close parallel with physiological transfer functions in midbrain network with a strong tuning to faster temporal modulations $\sim >$ 100 Hz and refined spectral tuning spreading up to ~ 4 cyc/oct. The tuning of individual model neurons is itself well-structured and localized in this spectro-temporal space with clear organization of subsets to a wide range of acoustic cues spanning frequency proximity, harmonicity, onset, and AM rate among others, as shown by the clustering analysis.

Traditionally, biomimetic computational models of stream segregation have attempted to replicate some or all of these cues to enable stream segregation. Often, this process is achieved by hand-selecting specific axes of feature analysis that best suit the auditory scenes of interest in these specific studies [29, 173, 174]. One of the drawbacks to feature selectivity in model design is confining the testable signals to those that take advantage of these specific features. By employing an unsupervised approach to feature selection, the current model not only replicates known simultaneous cues in auditory scene analysis, but also nonlinearly spans multitudes of features given the fully-connected nature of the Restricted Boltzman Machine (RBM) used in layer $\mathcal{L}_1$. Across-feature integration is in line with recent findings suggesting that many auditory neurons are driven by a multitude of stimulus features [175]. This feature integration is particularly crucial in case of complex sounds where a multitude of dimensions provide the perceptual system with converging evidence about the organization of the scene [115, 176]. The complementary value of this cross-feature mapping is clearly visible in control experiments where dropping different components of the simultaneous layer have different effects on the model's ability to perform stream segregation (Figure 3.3).

Role of Sequential Layer: Along the same lines, the sequential layer provides an integrated non-linear mapping of the feature space from localized details to slowly evolving spectro-temporal patterns. The use of a cRBM layer allows the model to 'learn' tuning from natural sounds along slower time-constants. The transfer function analysis reveals a strong selectivity to slow temporal modulations present in natural sounds typically in the range $\sim$ 2 - 32 Hz as shown in Figure 3.1B. This tuning is reminiscent of modulation transfer

functions derived from the mammalian auditory cortex with slightly broader neurons spectrally and slow temporally [147, 148, 167]. This global analysis has not been extensively investigated in models of auditory scene analysis, though few models have leveraged cortical-like processing to complement local feature analysis [29, 177–179]. Engineering approaches have also leveraged this global analysis especially in the case of speech processing systems. Approaches such as RASTA (relative spectra), high-pass and band-pass filtered modulation spectrum take advantage of slow articulatory structure of speech production as well as the sensitivity of human perception to such slow dynamics to offer a more robust processing of speech sounds in presence of noise [180–182].

Role of Temporal Coherence Layer: While the feature analysis is a crucial ingredient in auditory scene analysis, fusing the relevant cues together is an equally important complementary stage to group the features into meaningful objects. Perceptual and physiological data have strongly suggested that temporal coherence achieves the feature fusion needed for object formation [18, 34, 137, 183]. The current model employs biologically plausible Hebbian interactions across channels to rapidly adapting co-operative and competitive interactions between coherent and non-coherent responses [138]. Effectively, channels that exhibit a high degree of temporal correlation across feature dynamics are mutually strengthened while incoherent channels are gradually weakened hence facilitating segregation of target signals from background interference.

Scene segregation and fusion

A key contribution of this architecture is its ability to quantify the complementarity of rich feature representation and grouping mechanisms in driving scene segregation processes. The proposed architecture faithfully replicates human psychoacoustic behavior on steaming paradigms over wide range of stimuli ranging from simple tones to speech utterances as demonstrated in Figure 3.2. In case of two tone streaming paradigm shown in (Figure 3.2 A), the network exhibits stream segregation when two alternating tones are widely separated across tonotopic frequency axis. This behavior of the network is consistent with well established psychophysical and physiological findings of stream segregation induced by differences in tonotopic cues [184–187]. The primary reason for this behavior in proposed architecture is when ΔF is high in an alternating two tone sequence, two different groups of frequency selective neural units get activated in $\mathcal{L}_1$. In absence of temporal correlation between these two groups, the temporal coherence layer aided by adaptive mechanism suppresses the anti-correlated groups of units, hence inducing stream segregation in the final stage of the network. However when ΔF is small enough, there is high degree of overlapping between these two groups as a result of which the two tones get combined into a single stream.

When the two tone sequence is presented in a synchronous fashion, the network demonstrates strong fusion even when ΔF is high as shown in (Figure 3.2 A). The primary reason for this is when a synchronous two tone sequence evokes two different groups of co-active neural units, the temporal coherence layer strongly binds these two groups over time and allows the network in

fusing two tones into a single stream. This behavior strongly supports the spatio-temporal view of auditory stream segregation which requires neural channels to be widely separated as well as temporal incoherence across these channels [188]. It is also well consistent with the psychophysical findings suggesting that synchronous spectral components fuse perceptually into a single coherent sound, whereas any degree of asynchrony introduced across these components results in segregation [189].

Our network closely replicates the human psychoacoustic behavior in terms of demonstrating the effect of buildup over two tone stream segregation paradigm as shown in (Figure 3.2 B). The effect of buildup as established in [156, 190–192] suggests that stream segregation is a time dependent phenomenon in which the sequence of sounds are initially heard as a single stream, and that with time the same sounds split into two separate streams. The prime example of this behavior is demonstrated by our network for the case when frequency difference (ΔF) between the two tones is typically in the range $3-6$ semitones. Initially, the network starts with a single percept because of considerable overlap across the groups of frequency selective neural units. However, the temporal coherence layer (in presence of adaptive mechanism) makes sure that the binding of strongly correlated groups of units keeps on getting stronger and suppressing the anti-correlated units over time in the same process. This phenomenon improves the stream segregation performance of the network over time and demonstrates the buildup effect of stream segregation in a hierarchical framework.

Our network architecture exhibits similar stream segregation behavior in case of complex tones as shown in (Figure 3.2 [C,D,E]). This behavior is consistent

with past psychophysical findings which suggest that stream segregation is induced with sounds evoking segregated responses along any of the feature dimensions in the auditory pathway including fundamental frequency (F0), spectral and temporal modulation rate axes and onset among others [193–197]. The network performance validates our claim that localized spectro-temporal basis representation captured in $\mathcal{L}_1$ shows selectivity to different simultaneous cues like harmonicity, onset and AM among others. When the foreground and background sounds differ significantly with each other in terms of one of these cues, the temporal coherence layer in conjunction with adaptive mechanism makes sure that the co-active units responding strongly to the foreground are grouped together, whereas all the other competing units get suppressed, hence allowing the network to induce stream segregation at the output of final stage of hierarchy.

The performance of proposed network in correct identification of speech tokens (“number" and “color") over speech intelligibility paradigm further validates our claim that a combination of rich feature space and grouping mechanisms drive scene segregation processes in varied complexity of sounds ranging from simple tones to complex speech utterances. The findings are consistent with well established physiological and psychoacoustic theories of scene segregation which suggest that a rich span of feature space in auditory pathway underlies a distributed representation of natural scenes [114, 198, 199] whereas grouping mechanisms based on coherence of temporal structure provides an elegant solution to mapping of feature space to well defined object based representation in the auditory system [18, 137].

We have demonstrated in our control experimental results that tuning the

network by controlling the firing of certain $\mathcal{L}_1$ and $\mathcal{L}_2$ clusters significantly impacts the network performance on stream segregation and speech intelligibility paradigm. By switching off the $\mathcal{L}_1$ units showing high degree of selectivity to localized cues like harmonicity, onset and slow/fast temporal modulations as shown in (Figure 3.3) , the network is unable to process the information characterized by each of these cues and hence fails to map the acoustic representation to a discriminable feature space. Similarly, when $\mathcal{L}_2$ is just tuned to faster modulations, the network fails to phase-lock with the slower dynamics of the natural sounds present in the scene as a result of which the stream segregation and speech intelligibility performance gets severely impacted as shown in (Figure 3.4 - middle). We see a similar kind of effect on speech intelligibility and stream segregation performance when adaptation and temporal coherence layer is not incorporated in the network architecture as demonstrated in (Figure 3.4 - [left,right]) respectively. In absence of grouping mechanisms, the network fails to capture the mapping between feature space and sound percept, hence preventing the network from segregating non-coherent objects in the dynamic scene. The results of our control experiments validate the claim that a synergy of rich feature space and biologically driven grouping mechanisms drive the processes involved in segregation of constantly changing dynamic scene to coherent objects.

In next chapter, we extend this hierarchical framework to an event detection paradigm with an aim to explore bottom-up saliency mechanisms. Further, we also develop an acoustic scene classification based on this rich hierarchy of local and global attributes and use it to complement the bottom-up framework in the task of salient event detection.

Chapter 4

Abnormal Event Detection using hierarchy based bottom-up and top-down saliency

4.1 Introduction

Our surrounding soundscapes are constantly changing as we go about our lives; walking from an office to the street to a cafe and carrying conversations along the way. This constantly changing acoustic environment floods our auditory system with complex information which needs to be analyzed in order to make sense of the events around us. Human auditory system has an exceptional ability of sampling the surrounding environment to pay attention to the salient objects of interest, while ignoring irrelevant backgrounds Such an ability is guided by auditory attention, which is a process of allocating sensory and cognitive resources to objects of interest [200]. For instance, at a cocktail party, we can keep attending to someone's voice in a conversation and neglect the ambient background, however a shrill sound of telephone ring will cause us to shift our attention to the salient event.

Auditory attention can be conceptualized as a selection process which directs sensory-driven as well as cognitive mechanisms to focus its resources onto anything that is deemed salient in an incoming acoustic stimuli. The selection process is influenced by both “bottom-up" sensory driven factors as well as “top-down" task specific goals [201]. In case of purely bottom-up selection process, attention is triggered by a stimulus in which the sensory driven cues are salient enough for conspicuous events to pop up amidst surrounding sounds. These sensory driven cues are typically represented in terms of rich feature representation, believed to be the building block of an auditory object. The bottom-up saliency is characterized in terms of how these feature elements are distinct enough to enable their constituting object stand out in a scene amidst different objects. Bottom-up saliency is a major driving force behind attentional mechanisms, however in cases when an auditory object is not distinctively different among its neighbors in an acoustic scene, the attentional mechanisms are complimented with top down task defendant processes. The top-down task relevant process uses prior knowledge and learned past expertise to focus attention on the target locations in a scene. For example, in an acoustic mixture comprised of musical piece and spoken conversation, the attention of subject may shift to speech sound if task is "what is being spoken ?", while the attention may shift to the music if the task is "which instruments are being played ?"

Vision based saliency models have been very successful in leveraging bottom up and top down visual cues from complex scenes and use them to build well defined attention frameworks [202]. Bottom up visual saliency has been primarily captured as a distinction of image locations, regions or objects in terms

of low level cues such as color, intensity, orientation, shape, T-conjunctions, X-conjunctions, etc. [203]. Such low level features are shown to affect visual search and bias eye fixations in natural scenes [204]. The correlation between physical structure of visual scene and saliency based selective visual attention has been exploited via successful models based on spatial scales [205], local geometry [206], spectral contrast [207] using well defined statistical approaches like information entropy [208] and natural statistics [209]. Bottom-up saliency has been well complimented with a number of top-down approaches like bayesian framework exploiting cognitive features and scales for target identification [210], support vector machine (SVM) framework based on low, mid and high level visual cues for predicting human fixations [211] as well as to learn task driven object based visual attention [212]. In recent times, research efforts have shifted focus from use of hand-engineered features to use of data-driven features to augment saliency prediction using deep learning models such as convolutional neural network (CNN) [107, 213–215] and recurrent neural networks among others [216, 217]. Such saliency models based on pre-trained deep features have shown to be extremely effective in learning rich representational space of low level features [218] as well as more complex and high level embeddings capturing semantic information such as class of objects [219].

In auditory research, saliency driven computational models have a huge role to play in tasks like event detection and classification, audio tagging, audio summarization and audio surveillance among others [47, 220]. Most of event detection and classification studies in past decade have relied upon low-level features in capturing event specific characteristics. These features

include short-time energy (STE), zero-crossing rate (ZCR), linear predictive coding coefficients (LPC), periodicity and pitch information, as well as entropy distribution of discrete Fourier transform components [19, 20, 22, 71, 221]. Mel-Frequency Cepstral Coefficients (MFCC) have been the most widely used representation in acoustic event detection tasks because of its ability to capture compact and efficient mapping of spectral characteristics of simple events [23, 24]. However, as the acoustic scenes encountered in real life environments are more complex and dynamic in nature, it becomes imperative to capture both special and temporal nuances of the signal over multiple time-resolutions [28, 222]. In this regard, spectro-temporal features were introduced via employing two-dimensional time-frequency Gabor filter-banks, localized Fourier bases and even biomimetic spectro-temporal receptive fields to capture event modulation patterns in both time and frequency domains [72, 75]. Although, huge efforts have been made to designing optimal features, detection and classification of salient event in continuous audio still remains a challenging task because of limited scope of hand-crafted features in capturing salient characterization of an event in a complex acoustic environment.

In recent times, with the emergence of deep belief architectures, deep unsupervised representation learning techniques have achieved tremendous success in domains like music genre classification, phoneme classification and speaker identification [66, 117, 118, 120]. The key idea is to learn complex abstractions as rich feature space from natural soundscapes in a data driven fashion. Recently, unsupervised representation learning has begun to be applied to event detection and classification tasks with moderate success. In [223], deep belief networks (DBN) were employed to learn low level embeddings from

unlabeled data which were then fed into concatenated softmax layer for final classification. Cakir *et al.* in [64] used multi-label feed-forward deep neural networks (DNN) for polyphonic sound event detection and showed that with significant numbers of hidden layers, hidden units and training data, DNNs outperform conventional generative models like GMM and HMM in event detection and classification task. In recently conducted DCASE evaluations, it was shown that combination of CNN and LSTM framework is capable of extracting salient details from myriad of scenes which is reflected in boosted accuracy in detection and classification paradigms [224, 225]. Although intensive research effort is being continuously expended into developing deep net architecture to push forward the area of sound event detection and classification, most of these techniques are still limited to supervised setup which models each of the events specifically using a discriminative cost function and loads of augmented data. There is a lack of relevant studies that can explore the role of acoustic driven bottom-up saliency mechanisms in event detection paradigms as well as how to integrate bottom-up saliency with top-down task specific knowledge so as to bias the bottom-up resources onto locations of interest in a complex acoustic scene.

The current study proposes an architecture of salient event detection in a continuous audio using a combination of acoustic driven bottom-up saliency mechanism and top down task specific knowledge. The study is based on the hypothesis that sound evolves in a high dimensional feature space and a salient event is “flagged" whenever there is temporal irregularity across the feature representation [226, 227]. The study leverages recent advantages in deep learning to explore the emergence of acoustic cues across auditory

pathway in bottom-up fashion and how such cues adapt themselves to a particular scene by following the short and long term regularities of the scene [228]. We propose a multi-layered deep belief architecture that leverages the hierarchical and non-linear representation of sound along the auditory pathway. The network is trained in an unsupervised fashion on a rich sound dataset including speech, natural scenes etc. to learn a rich combination of *local* and *global* cues from natural soundscape. The deep belief architecture is comprised of two layers in which each layer is designed to extract different degree of details from input acoustic signal. The first layer is designed so as to learn a rich tiling of *localized* spectro-temporal bases from complex acoustic environment [66, 121]. The second layer is designed to perform a long term *global* analysis to aid the network in learning temporal regularities across *local* cues over multiple time-scales [124]. Finally, we incorporate an online feedback driven adaptation framework into the same architecture which adapts the spectro-temporal bases based on incoming statistics of an auditory scene. This adaptive framework tracks the temporal regularities of an incoming acoustic scene and flags any deviations from these regularities as “salient" [229, 230]. We also propose a top down scene classification framework to compliment the bottom-up saliency network in detection of events of “interest". The top down architecture operates on *local* and *global* cues learnt via deep belief network, further supplemented by three layer LSTM network exploiting the sequential representation of acoustic cues. We demonstrate two fold advantage of top down acoustic scene classification framework: 1) The framework exhibits comparable performance to state-of-the-art performance over acoustic scene classification task. 2) The bottom-up saliency network uses the top-down

event specific knowledge of classification framework to focus its attention on events of “interest" rather than any “new" event. We use this integrated framework of bottom-up and top-down driven saliency mechanism in an *abnormal* event detection paradigm and demonstrate how a bottom-up deep belief based network is capable of extracting salient characteristics of a complex acoustic scene and how predefined knowledge of *abnormal* events via top down classification framework guides the network in correct detection of events of “interest".

The organization of paper is as follows: Section 4.2 provides a detailed description of the proposed architecture. Section 4.3 outlines the experimental setup and network configuration and section 4.4 event detection and classification results. Section 4.5 provides a discussion of results in the context of role of bottom-up and top-down processing in driving saliency mechanisms in auditory scene analysis.

4.2 Proposed Architecture

The proposed architecture is structured along 5 stages of auditory processing as shown in Figure 4.1. The first stage of processing transforms an incoming acoustic signal to two dimensional time-frequency representation. This representation is the passed as an input to second stage which extracts localized spectro-temporal attributes from auditory spectrogram. The output of this stage is then passed as input to third stage which learns the temporal regularities across the local cues. The 4th stage performs short term and long term adaptation across multiple time scales based on temporal regularities

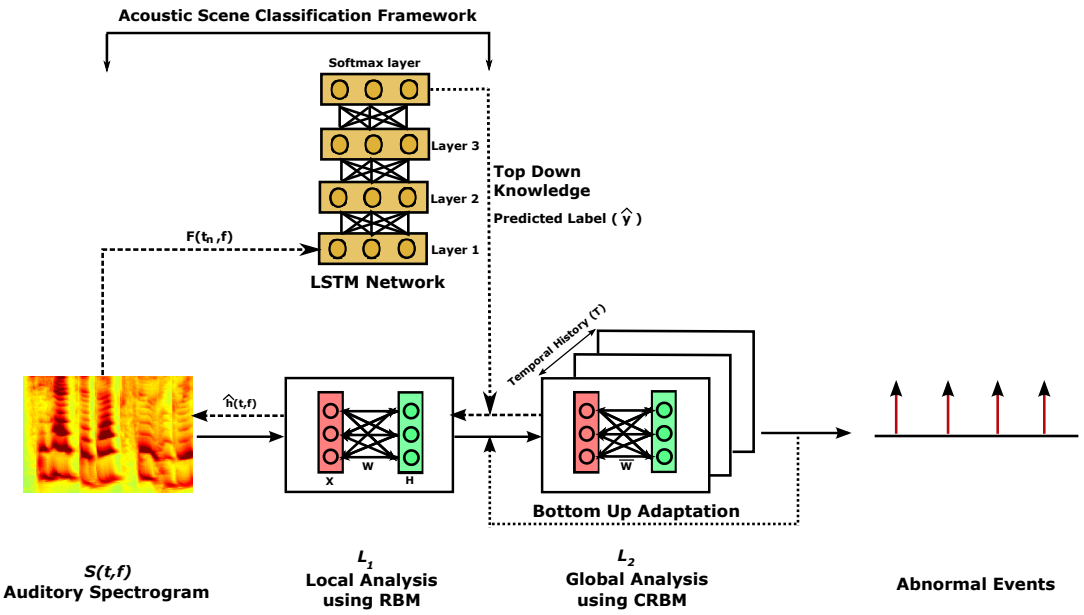

Figure 4.1: Schematic of proposed architecture. The architecture combines bottom-up and top-down processing for detecting abnormal events. Bottom-up processing is comprised of two deep belief layers performing local and global analysis and a multi scale adaptation framework. Top down processing is based on a scene classification paradigm passing down scene specific knowledge.

captured in previous stage. A top down classification framework is designed as the 5th stage in proposed architecture whose output is passed down as top down knowledge of specific event to compliment the bottom-up processing. Details of each stage are outlined next:

The acoustic signal is first processed through a model mimicking peripheral processing in mammalian auditory system [29]. The model transforms the acoustic signal into joint time-frequency representation referred to as auditory spectrogram. This stage starts with 128 symmetric filters equally spaced on a logarithmic axis over 5.3 octaves spanning the range 180-4000 Hz. Next, the filter outputs undergo spectral sharpening via first order derivative along the frequency axis followed by half wave rectification and short term integration

with $u(t,\tau) = \exp(-t/\tau)u(t)$ where $\tau = 10ms$. This analysis mimicking the cochlear analysis of Yang *et al.* [114] results in a time-frequency auditory spectrogram given by $S(t,f)$. 3 consecutive frames are then grouped together to form a one dimensional vector x in a process of shingling [133] such $x \in \mathcal{R}^n$ and $n = 384$. A dataset of n sampled patches given by $\mathcal{X} = x^1, x^2, \ldots, x^n$ is formed. This set of time frequency patches $\mathcal{X}$ forms the input to the next stage of processing.

4.2.1 Local Analysis using RBM

This stage is structured as Sparse Restricted Boltzmann Machine (RBM), chosen to discover features from unlabeled soundscape in an unsupervised fashion. Sparse RBM in this architecture is comprised of fully connected visible and hidden layer [231]. The visible layer units x_k are real valued and characterized by a Gaussian distribution fitted over input spectrogram $S(t,f)$; while the hidden units h_k are sampled from Bernoulli distribution such that $h_k \in 0,1$ for $k = 1,2,\ldots,N$, where N equals the number of nodes used in the hidden layer, which in this case is equal to 350. The network is parameterized by $\theta = W, b_x, B_h$ where W represents the interconnected weights between x and h, b_x represents the visible bias and b_h represents the hidden bias. The network is trained using Contrastive Divergence (CD) algorithm [134] so as to minimize the reconstruction error between x and $\hat{x} = hW + b_v$. We refer the reader to [111] for detailed understanding of the training paradigm. After the training, the connection weights W_k are transformed into two-dimensional

$h_k(t, f)$ where $t = 3$ and $f = 128$ to obtain a representation akin to spectro-temporal attributes. We apply these 2D filters over the time frequency patch $S(t, f)$ to obtain filter responses $r_k(t)$ given by:

$$\mathbf{r}_k(t) = \sum_f \int y_l(\tau, f) h(t - \tau, f) d\tau \tag{4.1}$$

The localized filter responses $r_k t$ form the feature representation for the next stage in the hierarchy as discussed below.

4.2.2 Global Analysis using CRBM

This stage $\mathcal{L}_2$ is structured to learn long term temporal regularities of the acoustic cues well characterized by the localized spectro-temporal filters in the previous component. A dynamical modeling approach is incorporated in this stage to capture the long term temporal regularities as well as the rate of such regularities. For a dynamical network to capture such regularities from filter responses $r_k(t)$, firstly, a range of temporal resolutions $\tau \sim$ (30 - 600 ms) is defined. Based on each temporal resolution window, instances of $r_k(t)$ are grouped in a process of shingling. The process is repeated for each window size and different sets of inputs $r_k^{\tau_1}(t), r_k^{\tau_2}(t), \ldots, r_k^{\tau_{\mathcal{N}}}(t)$ are created.

We use a conditional RBM (CRBM) methodology as our dynamical network which is a non-linear generative model for time series data that uses an undirected model with binary latent variables h, connected to a collection of visible units γ [136]. In our case, we are representing the distribution of visible units γ^{τ_n} by gaussian fitted over the filter responses $r_k^{\tau_n}(t)$ for $k = 1, 2, \ldots, K$ obtained from the previous component. At each time step t, we maintain

a history of last T time steps and store the visible variables corresponding to these time steps in a **history** vector referred to as $\gamma_T^{\tau_n}$. Each visible input $\gamma_t^{\tau_n}$ and hidden unit $h_t^{\tau_n}$ at a particular time step t and corresponding to a particular rate τ_n receives directed connections from $\gamma_T^{\tau_n}$ so as to capture the long term temporal dependencies across visible units. This dynamical model is defined by a joint distribution :

$$p(\gamma_t^{\tau_n}, h_t^{\tau_n} | \gamma_T^{\tau_n}) = \exp(-E(\gamma_t^{\tau_n}, h_t^{\tau_n} | \gamma_T^{\tau_n}))/Z(\gamma_T^{\tau_n}) \tag{4.2}$$

where $\gamma_t^{\tau_n}$ is a gaussian fitted representation of current filter response $r_k^{\tau_n}(t)$, $h_t^{\tau_n}$ is a collection of binary hidden units such that $h^{\tau_n} \in (0,1)$, $\gamma_T^{\tau_n}$ contains the history of past T filter responses, and Z is the partition function for proper normalization. The energy function E is given by:

$$\begin{aligned} E(\gamma_t^{\tau_n}, h_t^{\tau_n} | \gamma_T^{\tau_n}) = \frac{1}{2}\sum_i (\gamma_{it}^{\tau_n} - \hat{c}_{it}^{\tau_n})^2 - \sum_j h_{jt}^{\tau_n} \hat{d}_{jt}^{\tau_n} \\ - \sum_{i,j} \bar{W}_{ij}^{\tau_n} \gamma_{it}^{\tau_n} h_{jt}^{\tau_n} \end{aligned} \tag{4.3}$$

where $\bar{W}^{\tau_n}$ captures the interactions between the filter responses and hidden variables corresponding to each rate τ_n and the dynamical terms $\hat{c}_{it}^{\tau_n}$ and $\hat{d}_{jt}^{\tau_n}$ are linear functions of previous T filter responses $\gamma_T^{\tau_n}$, given by:

$$\hat{c}_{it}^{\tau_n} = \left(C_i^{\tau_n} + \sum_l A_{il}^{\tau_n} \gamma_{lN}^{\tau_n} \right) \quad \hat{d}_{jt}^{\tau_n} = \left(D_j^{\tau_n} + \sum_l B_{il}^{\tau_n} \gamma_{lN}^{\tau_n} \right) \tag{4.4}$$

where C^{τ_n} and D^{τ_n} are static biases and A^{τ_n} and B^{τ_n} are autoregressive model parameters. The dynamic biases $\hat{c}^{\tau_n}$ and $\hat{d}^{\tau_n}$ help in integrating the input over past T time steps and apply them as a bias to the visible unit $\gamma_t^{\tau_n}$ and hidden unit $h_t^{\tau_n}$ at current time step t. The parameter set $\theta = (\bar{W}, A, B, C, D)$ of CRBM

defined over 300 hidden units for each rate τ_n are learned using contrastive divergence (CD) approximation similar to RBM network. We refer the readers to [232] for detailed understanding of training paradigm of CRBM framework.

4.2.3 Bottom-up Adaptation

This stage is based on a feedback mechanism framework in which the filter outputs corresponding to each rate in $\mathcal{L}_2$ are adapted based on change in temporal regularities. The change in temporal regularities are characterized in terms of first order change in multi-rate filter outputs. Filter outputs $o_k^{\tau_n}(t)$ for $k = 1, 2, \ldots, 300$ and $\tau_n \sim$ (30 - 600 ms) are given by:

$$o_k^{\tau_n}(t) = \bar{W}_k^{\tau_n} \gamma_k^{\tau_n}(t) + \hat{c_k}^{\tau_n} \tag{4.5}$$

where $\bar{W}^{\tau_n}$ represents the connecting weights between hidden and visible layer, $\gamma^{\tau_n}(t)$ represents the gaussian fitted visible input representation of $\mathcal{L}_1$ filter responses $r^{\tau_n}(t)$ and $\hat{c}^{\tau_n}$ designates the dynamic visible biases learnt for each rate τ_n in $\mathcal{L}_2$. The filter weights $\bar{W}^{\tau_n}$ capture the temporal regularities across acoustic cues emanated from a scene. However, one major constraint imposed in this methodology is that $\bar{W}^{\tau_n}$ is time-independent which limits the ability of the network to adapt to changing regularities of an acoustic scene. In order to relax this constraint, we introduce a time dependent variable represented by $g_k^{\tau_n}(t)$ as linear **gain** component to account for the changing regularities. In order to compute $g_k^{\tau_n}(t)$, we take first order derivative of filter outputs $z_k^{\tau_n}(t)$ such that:

$$\Delta z_k^{\tau_n}(t) = \bar{W}_k^{\tau_n} \Delta \gamma_k^{\tau_n}(t) + g_k^{\tau_n}(t) \bar{W}_k^{\tau_n} \gamma_k^{\tau_n}(t) \tag{4.6}$$

After solving equation 6 for $g_t^{\tau_n}$, we get a closed form solution given by:

$$g_k^{\tau_n}(t) = (\Delta z_k^{\tau_n}(t) - \bar{W}_k \Delta \gamma_k^{\tau_n}(t))(\gamma_k^{\tau_n}(t)\bar{W}_k^{\tau_n})^{-1} \tag{4.7}$$

In order to formulate the feedback mechanism, firstly, we make sure that the time dependent gain $g_k^{\tau_n}(t)$ as computed in equation (6) is normalized such that $g_k^{\tau_n}(t) \in (0, 1)$, and secondly, $g_k^{\tau_n}(t)$ is projected onto $\mathcal{L}_2$ weights $\bar{W}^{\tau_n}$ in a recursive state space framework as suggested in [233, 234]. The state space formulation for adaptive weights are given by:

$$\bar{W}_k^{\tau_n}(t+1) = \beta g_k^{\tau_n}(t)\bar{W}_k^{\tau_n} + \zeta||\bar{W}_k^{\tau_n}||^2 \tag{4.8}$$

where β is given by $0.001/\tau_n$ which implies for large values of τ_n, $\mathcal{L}_2$ weights adapt slowly to the change in scene regularities whereas for small values of τ_n, β forces the $\mathcal{L}_2$ weights to adapt at a faster rate. This allows the network to track multiple rates of regularities in a scene, hence accounting for both slow as well as faster transitions. ζ is used as L2 regularizer to make sure that L2-norm of $\bar{W}^{\tau_n}(t) \in (0, 1)$. Bottom-up saliency mechanism is incorporated in the same framework by resetting the adaptation framework whenever $g_t^{\tau_n}$ exceeds a certain threshold M decided empirically. Hence, the mathematical formulation for bottom-up saliency framework as incorporated in this stage is given by:

$$\begin{aligned} \bar{W}_k^{\tau_n}(t+1) &= \zeta||\bar{W}_k^{\tau_n}||^2 && \text{if } g_t > M \\ \bar{W}_k^{\tau_n}(t+1) &= \beta g_k^{\tau_n}(t)\bar{W}_k^{\tau_n} + \zeta||\bar{W}_k^{\tau_n}||^2 && \text{if } g_t \leq M \end{aligned} \tag{4.9}$$

4.2.4 Top-down task specific knowledge

The role of this component in the proposed architecture is to pass down task specific knowledge to bottom-up saliency framework so as to allow it to focus its attention onto events of “interest". In order to gather sufficient knowledge about some specific events, this stage is structured as scene/event classification framework based on RBM-CRBM features supported by 3 layer LSTM as the backend classifier.

In this proposed classification framework, a deep belief RBM-CRBM based representation has been used as feature space. As studied in [66], RBMs typically capture localized spectro-temporal bases from acoustic input whereas CRBMs capture the long term temporal regularities which characterize the global attributes present in an acoustic scene. This feature presentation forms a part of bottom-up framework proposed in the architecture. We implement a top-down methodology of feature extraction in this framework to incorporate both the local and global attributes in the same feature space. Firstly, the $\mathcal{L}_2$ filter weights $\bar{W}^{\tau_n}$ are untangled into $\bar{W}^{\tau_n}_{n\text{x}K_1\text{x}K_2}$, where n represents number of slices tangled together in the shingling process, K_1 represents the number of hidden units in $\mathcal{L}_1$ network $(= 350)$ and K_2 represents number of hidden units $(= 300)$ in $\mathcal{L}_2$ network. The filter weights $\bar{W}^{\tau_n}_{n\text{x}K_1\text{x}K_2}$ corresponding to each rate τ_n are then projected onto $\mathcal{L}_1$ filter weights to obtain new set of weights given by: t $\hat{W}^{\tau_n}_{K_0\text{x}K_2} = \sum_n \left(W_{K_0\text{x}K_1}\right) \left(\bar{W}^{\tau_n}_{n\text{x}K_1\text{x}K_2}\right)$ where K_2 equals number of visible units $(= 384)$ in $\mathcal{L}_1$ network. These newly obtained set of weights $\hat{W}^{\tau_n}_{K_0\text{x}K_2}$ are then transformed into two-dimensional $\hat{h}^{\tau_n}_k(t, f)$ where $t = 3$ and $f = 128$ to obtain a spectro-temporal bases representation where $k = (1, 2, \ldots, K_2)$. The

final feature map is obtained by convolving each of these 2D spectro-temporal bases with input auditory spectrogram $S(t_n, f)$ such that the final feature map is represented by:

$$\mathcal{F}_k^{\tau_n}(t_n, f) = S(t_n, f) \circledast \hat{h}_k^{\tau_n}(t, f) \tag{4.10}$$

where t_n represents the time-steps in auditory spectrogram and $\circledast$ designates 'same' convolution.

In this top-down scene classification framework, a 3-layer LSTM network is used as the background DNN based classifier operating on top-down feature map representation as shown in Figure 4.1. LTSM layers are composed of recurrently connected memory blocks in which one memory cell contains three multiplicative gates. The gates perform continuous 'read', 'write' and 'forget' operations enabling the network to utilize the temporal information over a period of time. The hidden layers of LSTM network have self-recurrent weights which enable the cell in the memory block to retain previous information. We refer the reader to [235] for further details of LSTM network. In the proposed system, an acoustic image given by $\mathcal{F}_k^{\tau_n}(t_n, f)$ is used for sequential learning. Each acoustic image is of the form $\mathcal{F}_{f \mathrm{x} t_n}^{\tau_n}$ for k different filters. Hence, for a single utterance $u(t)$, we have $(k \mathrm{x} \tau_n)$ different versions of acoustic image $\mathcal{F}_{f \mathrm{x} t_n}$. We randomly sample k' images from the set of $(k \mathrm{x} \tau_n)$ images and use them as an input set for utterance $u(t)$. The same procedure is repeated for the entire set training set which is further used for training the LSTM network. This allows the network to learn multiple and distinct sequences of local and global details characterizing acoustic events. The output vector z_t is extracted

from input acoustic image $\mathcal{F}_{f \times t_n}$ through LSTM layers, which is further forwarded to softmax layer. Finally, the class probability $\hat{y}_t$ is predicted through the softmax layer.

In order to propagate the top-down event specific knowledge back to bottom-up saliency framework, segment-based class labels are obtained from the network. GIven z_t, the predicted class label for segment of duration τ_n is given by:

$$C_{segment}^{\tau_n} = \underset{i}{\text{argmax}} \sum_{t=1}^{T} log(P(\hat{y}_t = i|z_t) \tag{4.11}$$

where T represents the number of frames in segment of duration τ_n. For abnormal event detection task, we have a predefined set of abnormal events designated by $C_{abnormal} = (C_{ab_1}, C_{ab_2}, \ldots, C_{ab_N})$. The top down event specific knowledge from LSTM framework is incorporated into the bottom-up saliency framework by introducing a modification to equation (9) as given by:

$$\begin{aligned} \bar{W}_k^{\tau_n}(t+1) &= \zeta||\bar{W}^{\tau_n}||^2 \quad \text{if } (g_t > M) \;\&\; (C \in C_{abnormal}) \\ \bar{W}_k^{\tau_n}(t+1) &= \beta g_k^{\tau_n}(t)\bar{W}_k^{\tau_n} + \zeta||\bar{W}_k^{\tau_n}||^2 \quad \text{if } g_t \leq M \end{aligned} \tag{4.12}$$

4.2.5 Detection of Abnormal Events

The final stage in the architecture is structured as abnormal event detection framework. After combining the bottom-up acoustic driven representation and top-down task specific knowledge, L_2 filter outputs are extracted by projecting modified $\bar{W}_k^{\tau_n}(t)$ as obtained in equation (12) onto L_1 filter distribution $\gamma_k^{\tau_n}(t)$ for $k = 1, 2, \ldots, 300$ corresponding to each rate τ_n. Modified L_2 filter

outputs $\hat{o}_k^{\tau_n}(t)$ are given by:

$$\hat{o}_k^{\tau_n}(t) = \bar{W}_k^{\tau_n}(t)\gamma_k^{\tau_n}(t) + \hat{c_k}^{\tau_n} \tag{4.13}$$

In order to detect an '**onset**' of any abnormal event in complex acoustic scene, the 'adaptation' and 'reset' properties of $\hat{o}_k^{\tau_n}(t)$ are exploited. The reset points in $\hat{o}_k^{\tau_n}(t)$ signify an onset of any event of "interest" as suggested in equation (12). The reset points are determined by taking first order derivative of $\hat{o}_k^{\tau_n}(t)$ given by $\hat{o'}_k^{\tau_n}(t) = \Delta\hat{o}_k^{\tau_n}(t)$. The reset information is integrated by summing across all the filters and rates such that $\mathbf{o}(t) = \sum_{\tau_n}\sum_k \hat{o'}_k^{\tau_n}(t)$. The onset points are identified by extracting the locations of peaks in $\mathbf{o}(t)$. In order to assess the correspondence between events detected by proposed mechanism and actual events, the scenes are analyzed over overlapping bins of duration duration T_L. Results presented in this paper are for 3 second windows with a time step of 1 second. Each bin containing both the actual onset and detected onset is marked as a "hit'"; each bin containing only the detected onset is marked as "false". Based on these statistics, a receiver operating characteristic (ROC) curve is generated by varying the threshold based on strength of peaks corresponding to detected events.

4.3 Experimental Setup

4.3.1 Data

An ensemble of natural sounds comprising of speech segments from TIMIT speech database [236] and environmental sounds like ambient outdoor noises and animal vocalizations from BBC sound database [76] is used to train RBM

($\mathcal{L}_1$) and CRBM ($\mathcal{L}_2$) bases in our architecture in an unsupervised fashion. Speech dataset and from TIMIT comprises of male and female speakers and approximately amounts to 4 hours of data. BBC database has total of 2400 recordings, amounting to 68 hours of data. Examples of animal vocalizations in BBC include barking dogs, bleating goats and chattering monkeys. The ambient sounds include different types of environmental noises, for example, street, office, warfare and transportation among others. Speech utterances are approximately 3 seconds in length, while animal vocalizations are and ambient sounds are broken into 3 seconds and are windowed using a raised cosine window to avoid transient effects. All segments are down-sampled to 8 kHz and standardized to zero-mean and unit variance.

The dataset used for abnormal sound event detection framework contains audio recordings having abnormal events mixed with a plurality of background noises [237]. The proposed architecture for abnormal event detection is experimentally validated considering an application of 'audio surveillance' in which the set of abnormal events $C_{abnormal}$ is comprised of three classes of audio events: *scream*, *glass breaking* and *gun shot*. The dataset consists of 384 audio files, each of the files is about 3 minutes long and containing a sequence of 'normal' and 'abnormal' events laid on top of each other. The audio files are available at multiple SNR levels in which the normal and abnormal events are mixed at a specific value of SNR, SNR^p, with p={0 dB,5 dB,10 dB,15 dB, 20 dB, 25 dB, 30 dB}.

For the top-down classification framework, we have used the publicly available ESC-50 dataset [238]. The ESC-50 dataset consists of 2000 environmental recording, each of which are 5 seconds long. The recording are divided

into 50 equally balanced classes. These 50 classes are divided into 5 major groups, namely, animals, natural sounds, human non-speech sounds, interior/domestic noises and exterior/urban noises. The dataset contains “glass-break" as one of the classes in interior noise set, hence in order to incorporate the top down information of all the abnormal events used in our event detection analysis, we further add classes like “gun-shots” and “scream" to the database. The files are arranged in 5-folds for comparable cross-validation.

4.3.2 Network Configuration

The deep belief architecture formulated for acoustic driven bottom-up saliency is comprised of RBM and CRBM network in which both the networks are trained using the contrastive divergence (CD) approximation. RBM network is comprised of 350 hidden units and CRBM network corresponding to each temporal resolution (τ_n) consists of 300 hidden units. Both RBM and CRBM network are initialized with learning rate of $\lambda = 10^{-3}$ and a sparsity target of 0.5. The autoregressive weights in case of CRBM used a learning rate of $\lambda_A = 10^{-3}$ as used in [136]. A momentum term is also used in both the networks: 0.9 of the previous accumulated gradient was added to the current gradient. The training cases are presented to both the networks as “mini-batches" of size 500 and the weights are updated after each mini-batch.

For top-down acoustic scene classification framework, we used a LSTM network of 3 hidden layers, each with 200, 100 and 100 LSTM cells respectively. The network is trained over 300 epochs using cross-entropy error as the loss function supervised by one-hot encoding class vectors. The randomly ordered

mini-batches in each epoch is set to be 200. After a mini-batch is processed, the weights are updated using adadelta [239]. The input sequence is an acoustic image which consists of 500 frames of 128 dimensional auditory spectrogram frequency bins. The output layer consists of 52 softmax nodes identical to the number of acoustic scenes.

4.4 Experimental Results

For evaluating the performance of proposed architecture on abnormal event detection task, we considered two primary measures: correct detection of 'abnormal' event onsets termed as true positive rate (TPR), detection of abnormal onsets when only normal events are present termed as false positive rate (FPR). Figure 4.2 shows the ROC measure for proposed acoustic driven bottom-up salient event detection framework. The proposed architecture is compared against two baseline bottom-up feature driven methodologies. One baseline framework is based on MFCC based feature representation and the second is based on combination of 9 standard features namely *brightness*, *bandwidth*, *spectral flatness*, *spectral irregularity*, *pitch*, *harmonicity*, *temporal modulations*, *spectral modulations*, and *loudness*. Details of each of these features can be found in [240]. The performance measure of each of the techniques is based on the same onset detection paradigm, which is extraction of onset time stamps based on derivative of feature representation. Figure 4.2 clearly suggests that proposed framework based on adaptation and reset mechanism achieves the best performance in terms of true positive rate. For an average false positive rate of 30% on the entire test set, the proposed framework achieves an average

correct detection rate of 71.2%. We consider the area under the ROC curves (AUC), which is equal to 1 for a perfect classification, as a measure of the performance of three feature driven bottom-up frameworks. The higher this measure, the better the overall performance of the framework is. The average AUC measure for the proposed bottom-up framework is 0.79 which is 24% relative improvement over the performance of MFCC (AUC=0.6) and standard feature set (AUC=0.55).

To evaluate the performance of proposed top-down scene classification

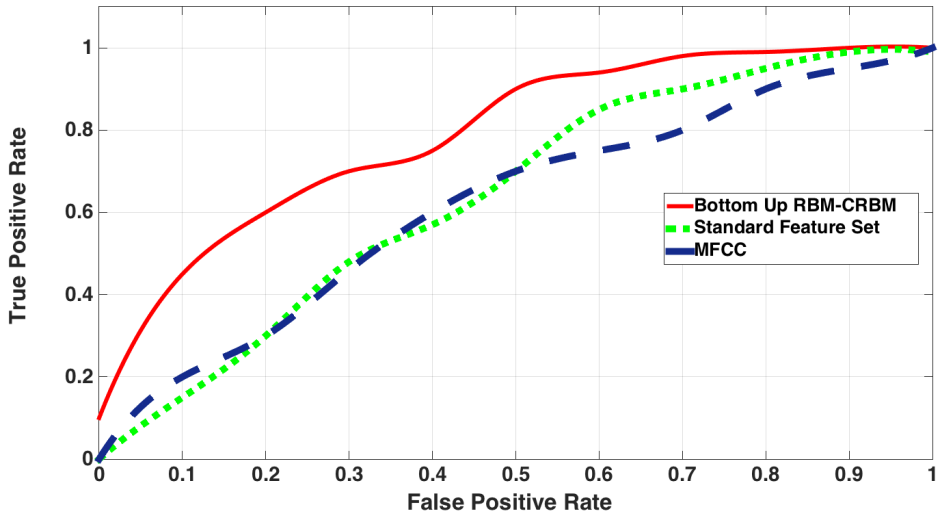

Figure 4.2: ROC measure for three bottom-up feature driven configurations related to abnormal sound event detection

framework, 5-fold cross validation was performed on 52 classes (ESC-50 + “gun-shot" + “scream"). The classification accuracy for each class of scene is reported in Figure 4.3 for the proposed RBM-CRBM-LSTM framework and compared against a framework using the same set of feature representation extracted from RBM-CRBM network supported by a GMM based classifier

in the backend [81]. Figure 4.3 shows that the proposed architecture exhibits an improved accuracy across majority of classes compared to GMM framework. However, such improvement is not noted across *all* classes, hence suggesting that sequential modeling of *local* and *global* attributes captured via RBM-CRBM is not enough for discriminable representation of a complex acoustic scene. We further compare our framework's performance against other studies in literature reported on same task and summarize the results in Table 4.1. The comparison suggests that the average classification accuracy of 87.2% achieved by our proposed framework is slightly better than state-of-the-art network of Filter Bank Energies (FBEs) and convolutional RBM based feature representation and CNN based classifier as reported in [223].

The top down knowledge acquired from scene classification is further

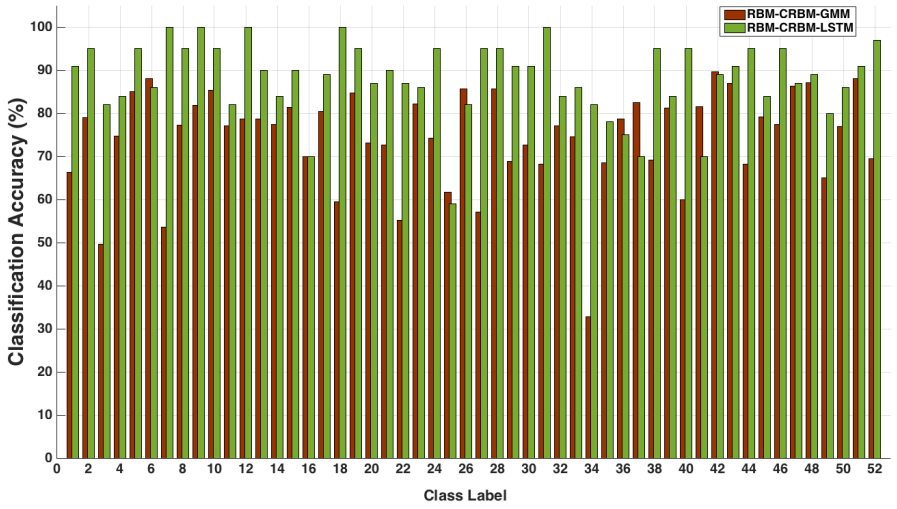

Figure 4.3: Classification Accuracy for each of the fifty-two classes for the proposed and baseline generative framework

incorporated into bottom-up saliency framework to augment the abnormal

Framework	Accuracy(%)
RBM-CRBM-LSTM	**87.2**
FBEs-ConvRBM-Bank [223]	86.5
Piczak FBEs-CNN [241]	64.5
EnvNET [242]	64.0
logmel-CNN [242]	66.5
logmel-CNN-EnvNet [242]	71.0

Table 4.1: Comparison of classification accuracy of ESC-50 dataset in literature

event detection paradigm as per the schematic shown in Figure 4.4. Figure 4.4 suggests that the top down classification framework generates labels for each segment 1 second long. If these labels don't match a corresponding abnormal event, adaption-reset mechanism is driven by pure bottom-up saliency. However, if the labels generated by top-down framework matches an abnormal event and the bottom-up saliency misses the onset owing to weak saliency in changes across regularities, the top-down decision forces the bottom-up framework to reset its resources and restart the process of adaptation based on new event regularities. Hence, the incorporation of top-down knowledge into bottom-up framework augments the network in detecting the onsets of abnormal events which might get missed in the pure bottom-up process.

Figure 4.5 shows the ROC performance curve for the proposed top-down and bottom-up based saliency framework for abnormal event detection. The proposed framework is compared against baseline architecture proposed by Foggio et.al [237] for the same dataset. The top down knowledge of abnormal classes is gathered from scene classification framework developed on dataset comprised of ESC-50 scenes and "gun-shot" and "screams", hence, in order to provide a fair comparison, we replicate the scene classification framework

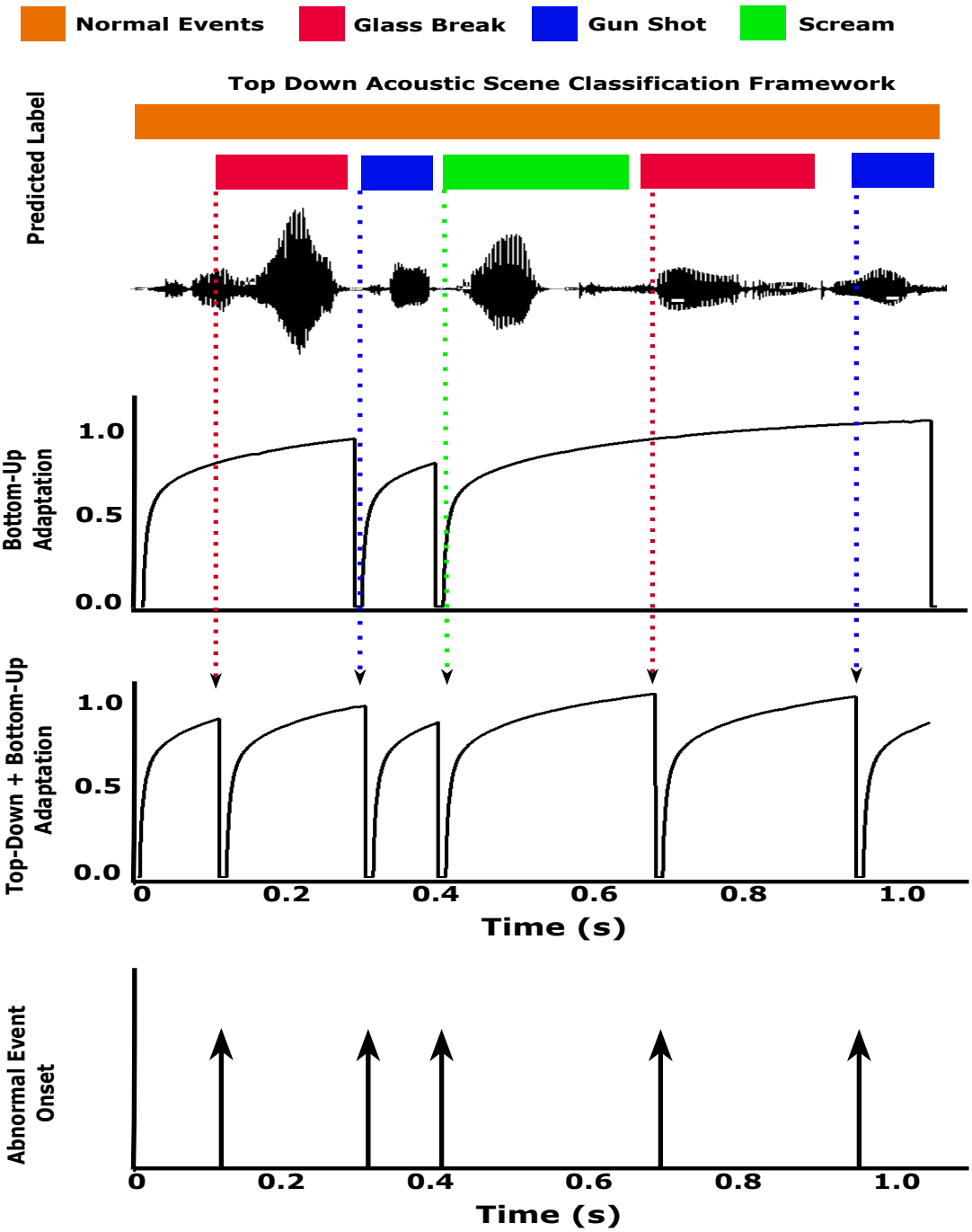

Figure 4.4: A schematic of top-down event specific knowledge incorporated into bottom-up framework

using the baseline architecture as proposed in [237] for the dataset comprised of ESC-50 scenes and “gun-shot" and “screams", and subsequently use it for abnormal event detection. The baseline architecture is a pure top-down based event detection framework which consists of short-time and long-time descriptors in terms of feature representation supported by SVM based classifier. We refer the reader to [237] for detailed description of feature set and learning mechanism. Figure 4.5 clearly suggests that incorporation of top-down task specific knowledge of abnormal events like “glass-break", “gun-shot" and “screams" into bottom-up framework improves the performance of the architecture in terms of true positive rate. The combination of top-down and bottom-up based saliency framework achieves an average correct detection rate of 78.8% for an average of false positive rate of 30% which is 7.6% absolute improvement over performance of proposed bottom-up framework as shown in Figure 4.2. The proposed framework also achieves almost 19% absolute improvement over baseline SVM based architecture, hence signifying the importance of bottom-up processes in driving saliency mechanisms. For further analysis, AUC is computed from ROC curves of each of the frameworks shown in Figure 4.5. The top-down$\oplus$bottom-up architecture achieves an AUC measure of 0.84 compared to 0.79 for bottom-up framework and 0.65 for baseline architecture. This measure further validates the fact that incorporation of top down knowledge of abnormal scenes in saliency mechanism compliments the bottom-up saliency driven mechanism and significantly improves the performance of network in detection of abnormal events.

To test the robustness of the proposed framework, a comparative analysis of bottom-up and top-down$\oplus$bottom-up architecture is performed for the

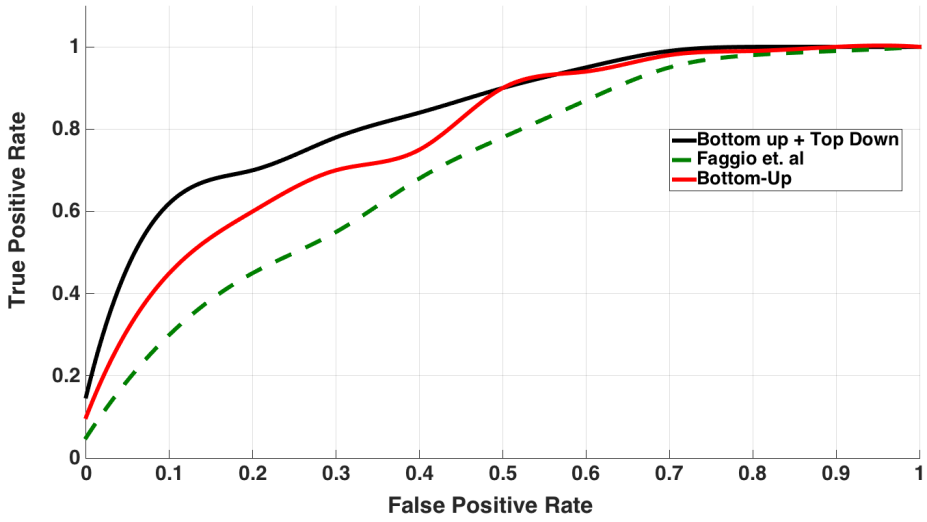

Figure 4.5: Comparison of ROC measure for proposed top-down⊕bottom-up, pure bottom-up and SVM based baseline architecture in abnormal event detection test-set

entire test set at different levels of SNR. In the context of abnormal events mixed with normal events in a continuous audio, high SNR means the abnormal events are more prominent in the audio. As the SNR level decreases, the prominence of normal events in the audio increases which subsequently increases the complexity of the abnormal event detection task. The results for this analysis are reported in Table 4.2. The performance of both the architectures are reported in terms of F1 and ER (error-rate) score. Table clearly suggests that top-down⊕bottom-up architecture matches the performance of pure acoustic driven bottom-up architecture at high SNR and exhibits high degree of robustness at low SNR. When normal events tend to get more and more prominent in the audio in terms of low SNR level, it becomes difficult for bottom-up architecture to detect the onsets of abnormal events, as reflected by low F1 and high ER score.

SNR	F1 score		ER score	
	Bottom-Up	Bottom-up $\oplus$ Top-down	Bottom-Up	Bottom-up $\oplus$ Top-down
30 dB	0.95	**0.96**	0.18	**0.16**
25 dB	0.92	**0.93**	0.20	**0.19**
20 dB	0.9	**0.91**	0.22	**0.21**
15 dB	0.83	**0.89**	0.28	**0.24**
10 dB	0.75	**0.85**	0.34	**0.27**
5 dB	0.65	**0.83**	0.41	**0.29**
0 dB	0.55	**0.71**	0.45	**0.30**

Table 4.2: Abnormal sound events detection results for proposed configurations at different SNR levels

4.5 Discussion

This study presents a hierarchical framework of salient event detection based on acoustic driven bottom-up mechanism and top-down task specific knowledge. The hypothesis for this framework is acoustic driven bottom-up saliency mechanisms are heavily dependent on a rich feature representation that can characterize an event present in a complex acoustic scene in terms of its specific details and can pop the event out in a neighborhood of other events based on how distinct and discriminative the details are. The hierarchical architecture is based on an unsupervised framework of learning the *local* and *global* attributes from a complex acoustic scene via a deep belief based **RBM-CRBM** network. This network is primarily drive by methodology which encodes two-dimensional input spectrogram into localized spectro-temporal basis representation via short term feature analysis as well as learning the long term temporal regularities across such bases. Based on this hierarchy based auditory representation, a bottom-up saliency mechanism is developed

based on *adaptation* of spectro-temporal bases as per change in regularities of complex heterogeneous event in a acoustic scene as well as *reset* of the same bases whenever a “new" event is salient enough to draw the attention of bottom-up framework. The study also proposes a top-down acoustic scene classification framework which exploits the sequential representation of *local* and *global* attributes via deep LSTM networks. The top-down event specific knowledge gathered via this framework is then used to bias the bottom-up resources towards the events of “interest".

A lot of studies in literature have explored the role of various features in the context of bottom-up saliency mechanism. Often, this studies are based on hand-selecting specific sets of features best suiting acoustic events of interest used in the studies [243–245]. One major limitation to feature selectivity in model design is the span of acoustic events which are limited to those which can take advantage of specific sets of features in saliency studies. By employing an unsupervised framework of feature learning, the proposed network is capable of learning a rich non-linear space of local and global attributes in a complex acoustic scene. This non-linear feature space is particularly crucial in case of complex scenes in which a multitude of dimensions provided converging and complimentary evidence about the salient organization of events in a scene. The characterization of long term temporal details via global analysis (CRBM framework in this study) has not been extensively studied in saliency models, though few models have explored the idea of complimenting local feature space with long term details, especially in speech processing systems which incorporate feature dynamics mimicking the articulatory structure of speech production to impart robustness to such systems in presence of noise

[180, 181].

A key contribution of the proposed architecture is a pure acoustic driven bottom-up saliency framework used in the context of abnormal event detection. Figure 4.2 shows that proposed bottom-up framework performs significantly better than two baseline frameworks used in previous studies [47, 246] in terms of correct detection of abnormal events as well as AUC measure. One key reason behind such an improvement is the complimentary information imparted by local and global specto-temporal bases learnt via RBM-CRBM framework related to salient characterization of fast as well as slow changing events in a complex acoustic scene. Another key reason is that the framework takes into account the change in regularities of the events in a continuous audio in terms of first and second order statistics of filter outputs in $\mathcal{L}_2$ and adapt the filter weights based on these statistics. The reset mechanism allows the network to bias its computational resources towards a salient change in event regularities, hence aiding the network in salient event detection. The normal events present in the test-set for the event detection task are typically comprised of household ambient sounds including combination of people conversing, telephone ring, cooking utensils and sound of water gushing out of tap among others, rain sounds, street ambient sounds etc. When normal events tend to get more prominent at low SNR level in a continuous audio, the saliency doesn't remain confined to abnormal events owing to heterogeneity of complex events. For example, at low SNRs, sound of telephone ring in the middle of a conversation will be detected as salient event in the proposed bottom-up framework, hence contributing to false alarms in terms of abnormal event detection. To tackle this constraint of bottom-up framework, the

top-down knowledge of abnormal events via an acoustic scene classification framework is incorporated in the same framework which allows the network to bias its bottom-up resources towards a salient event only when the saliency of the event is in agreement with top down event specific knowledge as shown in Figure 4.4.

A major contribution of this work is a top-down acoustic scene classification framework exhibiting comparable performance to state-of-the art [223] on ESC-50 dataset as shown in Table 4.1. The rich complimentary space of local and global spectro-temporal bases is driving the feature representation stage via RBM-CRBM framework which is further supplemented by 3 layer LSTM network exploiting the sequential representation of the feature space. The fact that the proposed framework betters the state-of-the art validates the importance of global analysis complimenting local feature space. However, for broadband classes like “rain" and “water-drops", the classification accuracy of the framework is well below mean classification accuracy of 87.2% as shown in Figure 4.3. This suggests that proposed feature space is not enough to characterize broad-band spectro-temporal representation. The proposed framework exhibits a classification accuracy of 94.8% for “glass-break", 90.1% for “scream" and 97.2% for “gun-shot". This analysis forms the basis for incorporating top-down information of such classes forming the core of abnormal events into the bottom-up framework.

The feedback of top down information in terms of posterior probability of an acoustic class into bottom-up framework lead to an improved performance in abnormal event detection task over pure bottom-up framework as well as SVM based top-down event detection framework [237] as shown in Figure 4.5.

The major factor accounting for such an improvement is that whenever an abnormal event $\in$ {glass-break, gun-shot, scream} shows up in the continuous audio, the top-down scene classification framework produces the right label $\hat{y}$ corresponding to such events which in turn forces the bottom-up framework to bias its saliency resources towards such an event. Table 4.2suggests that even at low SNR when the normal events are way more prominent than abnormal events, the proposed bottom-up$\oplus$top-down framework exhibits a F1-score of around 0.71 compared to 0.55 for pure bottom-up. This shows that the top-down classification framework is able to produce the right labels for abnormal events event at 0 dB SNR owing to its rich feature representation and deep LSTM network, which in turn imparts high degree of robustness into event detection paradigm. The ER score of 0.30 primarily accounts for the fact when the normal events are spanned by events like “rain sounds" and others which typically have a broadband spepctro-temporal representation, the feature space of local and global attributes fail to capture the salient changes in temporal regularities and hence, miss the salient onsets of abnormal events. Another major reason is at low SNR, an abnormal event like “scream" is confused with classes like “crying baby" or "laughing" which prevents the top-down network from feeding back the information of “scream" into bottom-up framework and hence, lead to high miss rate at low SNR.

Chapter 5

Abnormal Event Detection using Mixtures of Temporal Trajectories

5.1 Introduction

As discussed in Chapter 4, abnormal event detection in a continuous audio depends to a a great extent over salient details captured by the feature representation which makes a salient event "pop" out in the context of the other events in the neighborhood. However, when the events tend to get very unstructured and broad in nature, defining 'abnormal' behavior in an audio recording becomes a challenging task. First of all, there is no universal definition of what *abnormality* means. Second, even what is *normal* cannot be easily defined given the complex nature of sound sources in realistic scenarios. To date, most research efforts in anomaly detection have mainly focused on detection of isolated events in continuous recordings such as shouts [247], screams [248], laughs [249], gunshots and explosions [250] etc. However, for setting up a surveillance system in an environment like a train or subway station, detecting abnormalities based on examining isolated events becomes

highly inefficient since collections of such isolated events can overlay normal behavior. Instead, we consider the problem of obtaining a good model representation of normal behavior in the environment. We are particularly interested in models that can capture non-trivial commonalities across various sound events as well as their interactions in the context of a complex scene.

Modeling acoustic scene behavior ultimately reduces to a choice of feature representation and learning model that can best characterize the myriad events that can be encountered in acoustic scenes. Mel-Frequency Cepstral Coefficients (MFCC) are the most widely used representation in acoustic event detection tasks. They provide a compact and efficient mapping of the spectral characteristics of simple scenes [23, 24, 73]. Unfortunately, their performance does not generalize to real world environments which are inherently dynamic and often corrupted by noise. In order to accurately report the intricacies of such realistic scenarios, it is imperative that any modeling of acoustic characteristics captures both spectral and temporal nuances of the signal over multiple resolutions and time-constants [28, 222]. Work in this direction has often employed two-dimensional time-frequency filter-banks using Gabor filters, localized Fourier bases or even biomimetic spectro-temporal receptive fields [72]. In [66], Lee *et al.* reported a localized and rich tiling of the spectro-temporal space of sound classes derived from unsupervised learning over unlabeled data in the context of *Restricted Boltzmann Machines* (RBM) [251]. In the current work, we build on this rich basis set; and extend applicability of unsupervised learning using RBMs to the problem of anomaly detection in audio recordings.

Operating on this feature analysis often comes a robust backend classifier

whose role is to capture variability across different instances of the sound class. Unsupervised classifiers like Support vector machines (SVM) and Gaussian mixture models (GMM) have proved to be very efficient in modeling the mean statistics of analytical audio features in tasks of scream, laughter and gunshot detection [252, 253]. These models do provide well defined average representation of isolated events but fail to capture the information contained in the temporal dynamics of these events. In contrast, HMM based models are capable of capturing such temporal trajectories [253]. However, because of their markovian constraint, they become inefficient in modeling the long term temporal dependencies across events essential to obtain a global context of an acoustic scene. Recent work started using more representationally powerful generative models based on distributed hidden states, such as Conditional RBMs [254] to learn representation of temporal dynamics from data rather than explicitly modeling them under hard wired assumptions.

In the current study, we develop a hybrid RBM-CRBM scheme for modeling normal acoustic behavior in a subway station. An “event" such as normal conversation among riders is typically comprised of multiple sub-events like speech, laugh, cheerful banter etc., each having its own set of spectral and temporal dynamics. In order to capture these different modes of temporal dynamics as well as their interactions and transitions across each other, we propose a *mixture of dynamic trajectories* that can decompose the global temporal space of a normal event into multiple trajectories, each of which belongs to a semantically different sub-event. We develop an integrated framework of learning the localized spectro-temporal attributes in an unsupervised fashion as well as capturing their different modes of temporal trajectories by using a

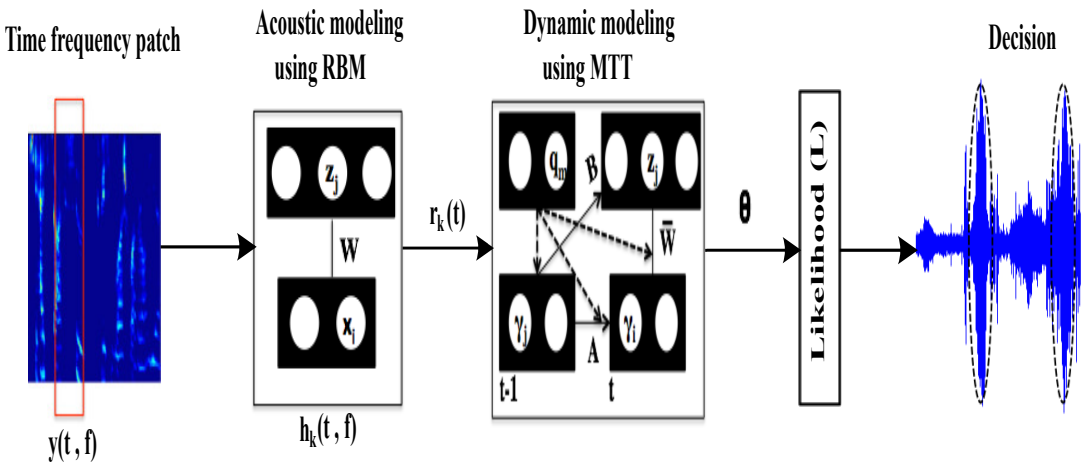

Figure 5.1: Block diagram of MTT based abnormal sound event detection

set of mixtures of temporal trajectories (MTT). The framework flags anything as an '*abnormal*' event that don't fall within the span of learned trajectories. The organization of this paper is as follows: Section 5.2 provides a detailed description of the proposed methodology using a hybrid RBM-MTT framework. Section 5.3 outlies the experimental setup and event detection results, while section 5.4 provides conclusion and discussion of the results.

5.2 Method

Our proposed framework for abnormal sound event detection comprises 3 main processing blocks; acoustic modeling using RBM, dynamic modeling using MTT and finally using these models for abnormal sound event detection as shown in Figure 5.1. The system operates on time frequency representation of acoustic signals. A time-frequency auditory spectrogram $y(t, f)$ is extracted from each audio file based on a model of peripheral processing in the mammalian auditory system [29]. The spectrogram representation $y(t, f)$

is sampled with frame size of 10 ms. 10 consecutive frames are then grouped together to form a one dimensional vector x in a process of shingling [133]. A dataset of n sampled patches given by $X = x^1, x^2, \ldots, x^n$ is formed, where $x^{(i)} \in R^N$ and N = 1280 in our case.

5.2.1 Acoustic modeling using RBM

We use Sparse restricted Boltzmann machine (RBM) as the unsupervised learning algorithm to discover features from the unlabeled dataset X. Sparse RBMs are undirected graphical models with K binary hidden variables [255]. We train the first layer RBM representations comprised of 400 hidden units using the contrastive divergence (CD) approximation with same type of hyper-parameters and sparsity penalty as used in [256]. The training produces the weights W_k for $k = 1, 2, \ldots, 400$ which are a representation of localized spectro-temporal attributes. In order to get a representation similar to localized 2D filters, we transform these one dimensional weights W_k into $h_k(t, f)$ where t = 10 and f = 128. We apply these 2D filters over the time-frequency patch $y(t, f)$ extracted from the labeled dataset of normal conversations to obtain filter responses $\mathbf{r}_k(t)$ given by:

$$\mathbf{r}_k(t) = \sum_f \int y_l(\tau, f) h(t - \tau, f) d\tau \tag{5.1}$$

Filter responses $\mathbf{r}_k(t)$ are used as our feature representation of 'normal' events for the next processing block.

5.2.2 Dynamic modeling using MTT

Next, a mixture of CRBMs (mCRBM) [8] is proposed as a *dynamical* mixture model to decompose the global temporal space of a normal event into multiple trajectories, where each such trajectory belongs to a particular sub-event. A *dynamical* mixture model can be created by introducing a mixture component variable, $\boldsymbol{q}$, with M possible states. The dynamical model is defined by a joint distribution:

$$p(\boldsymbol{\gamma}_t, \boldsymbol{z}_t, \boldsymbol{q}_t | \boldsymbol{\gamma}_N) = exp(-E(\boldsymbol{\gamma}_t, \boldsymbol{z}_t, \boldsymbol{q}_t | \boldsymbol{\gamma}_N))/Z(\boldsymbol{\gamma}_N) \tag{5.2}$$

where $\boldsymbol{\gamma}_t$ is real valued representation of current filter response $\mathbf{r}_k(t)$, $\boldsymbol{z}_t$ is a collection of binary hidden units such that $z \in (0,1)$, and $\boldsymbol{\gamma}_N$ contains the history of past N filter responses to provide a way for capturing the long term temporal dependencies across the responses. The energy function $\boldsymbol{E}$ is given by:

$$\begin{aligned} E(\boldsymbol{\gamma}_t, \boldsymbol{z}_t, \boldsymbol{q}_t | \boldsymbol{\gamma}_N) = \frac{1}{2}\sum_i (\boldsymbol{\gamma}_{it} - \hat{c}_{it})^2 - \sum_j \boldsymbol{z}_{jt}\hat{d}_{jt} \\ - \sum_m \boldsymbol{q}_{mt} \sum_{i,j} \bar{W}_{ij}\boldsymbol{\gamma}_{it}\boldsymbol{z}_{jt} \end{aligned} \tag{5.3}$$

where $\bar{W}$ captures the interactions between the filter responses and hidden variables and the dynamical terms $\hat{c}_{it}$ and $\hat{d}_{jt}$ are linear functions of previous N filter responses $\boldsymbol{\gamma}_N$, given by:

$$\begin{aligned} \hat{c}_{it} &= \sum_m \boldsymbol{q}_{mt} \left(C_{im} + \sum_l A_{ilm}\boldsymbol{\gamma}_{lN} \right) \\ \hat{d}_{jt} &= \sum_m \boldsymbol{q}_{mt} \left(D_{jm} + \sum_l B_{ilm}\boldsymbol{\gamma}_{lN} \right) \end{aligned} \tag{5.4}$$

where C and D are static biases and A and B are autoregressive model parameters. The parameter set $\theta = (\bar{W}, A, B, C, D)$ of mCRBM are learned using contrastive divergence (CD) approximation. We refer the reader to [8] for details of learning mCRBM by CD. This learned parameter set θ becomes our representation of mixture of temporal trajectories (MTT) models. We use M=10 assuming a mixture of 10 components can span the entire temporal trajectory space of a single event and use 200 hidden units in our mCRBM architecture.

5.2.3 Abnormal Sound Event Detection

In the detection stage, we use the measure of log-likelihood score of a given test frame under our learned MTT model to decide whether the frame under consideration belongs to an abnormal event or normal conversation [8]. A test audio signal is processed through the learned RBM weights $h_k(t, f)$ to obtain feature representation $\mathbf{r}_k(t)$ as per equation 1. On applying the parameter set θ over $\mathbf{r}_k(t)$, we obtain a log likelihood score L given by:

$$L = log(p(\mathbf{r}_t|\mathbf{r}_N; \theta)) = log(\sum_{\mathbf{z}_t, \mathbf{q}_t} p(\mathbf{r}_t, \mathbf{z}_t, \mathbf{q}_t|\mathbf{r}_N)) \tag{5.5}$$

We compare this likelihood score L with a threshold value obtained from development set and we label the frame as 'normal' if $L >$ threshold or 'abnormal' if $L <$ threshold.

5.3 Experimental Setup and Results

5.3.1 Data

We prepare an unlabeled training dataset by randomly mixing the recordings from both TIMIT [236] and BBC sound effects library [76] to train our first layer RBM bases. BBC sound effects library contain classes like Ambience, Animals, Office, Transportation and Musical etc. Because of such heterogeneity across the scenes, RBM weights are not biased towards one particular kind of scene. The dataset used for abnormal sound events detection contains recordings of audio events in a metro station [257]; the duration of each file ranging from 1 minute to about 6 minutes. We resample each recording in the dataset to 8 KHz and preprocess them through a pre-emphasis filter with coefficients $[1 - 0.97]$ in order to boost the high frequencies. The recordings contain events like normal speech, laughter, cheerful banter etc. annotated as *normal conversation*. The frames belonging to normal events are split randomly into 80 % for training the MTT models and rest 20 % as development and test set. The recordings also contain events like train passing by, shout, scream, fights, aggressive behavior etc. which we consider as 'abnormal' in our analysis and include them in the test set for detection.

5.3.2 System variants

The performance of an abnormal sound event detection system depends on how good our model representation is. The key aspect of our model representation is based on a set of mixtures of temporal trajectories capturing

the interactions and transitions across multiple events in a complex acoustic scene. In order to quantify its importance and effect on system performance, we contrast our proposed system against 3 system variants based on similar generative framework and backbone architecture but with variabilities in mixture components and trajectory representations. In one case, we train our MTT model using $M = 1$ to see how the performance of the detection system changes when a single mixture component is used to model different modes of temporal trajectories existing within a single event. Secondly, in order to quantify the importance of temporal trajectories based representation, we build a detection system by replacing mCRBM block with a regular RBM that models only the localized spectro-temporal modulations without any information of long term temporal dependencies. Our final system is based on learning first layer RBM bases only from normal conversation and use these learned bases for detecting the abnormal events.

5.3.3 Results and Analysis

Figure 5.2 shows the ROC for each of the detection systems by including/excluding the MTT stage as well as varying the number of mixtures capturing the temporal trajectories. The figure shows that our proposed system using MTT model with $M = 10$ performs the best in terms of true positive rate. When MTT is replaced by a RBM layer in the framework, we see that the detection performance of the system degrades because of incapability of RBM based representation in capturing the long term temporal dependencies. Single layer RBMs trained only on normal conversations gives the worst performance

in terms of true positives; the main reason being the first layer RBM bases trained on a small set of data are not able to capture a good representation of localized spectro-temporal attributes. As a result of this poor characterization, we get a lot more false negatives for this system compared to other systems. For MTT model with $M = 1$, the observation gets interesting. We see that its detection performance is better than RBM based systems, thus illustrating the importance of long term temporal dependencies over short term temporal structure for better characterization of sound events. However, the true positive rate for this system decreases when compared to MTT model with $M = 10$. This observation is mainly accounted for by the fact that due to presence of different modes of temporal trajectories within an event of normal conversation, MTT with $M = 1$ fails to span the entire temporal trajectory space of such a broader class. As a result, when an event like *laughter* occurs in a continuous audio, the system detects it as an abnormal event even though it is labeled as *normal conversation*.

To provide more insight into the idea of MTTs capturing different modes of temporal trajectories, we apply our MTT model to a sample recording of normal conversation among riders in a subway station. At several points during the conversation, other than normal speech, there are instances of *laugh*, *excitepment* etc. which are non stationary events having their own set of dynamics. As suggested in Figure 5.3, we find that frames belonging to the instances of *laugh* and *excitepment* are assigned to components 1, 4 and 7 with an average probability of 0.9572; while component 9 captures the temporal trajectories of normal speech in the conversation with an average probability of 0.9851. This probabilistic assignment of frames to different components of

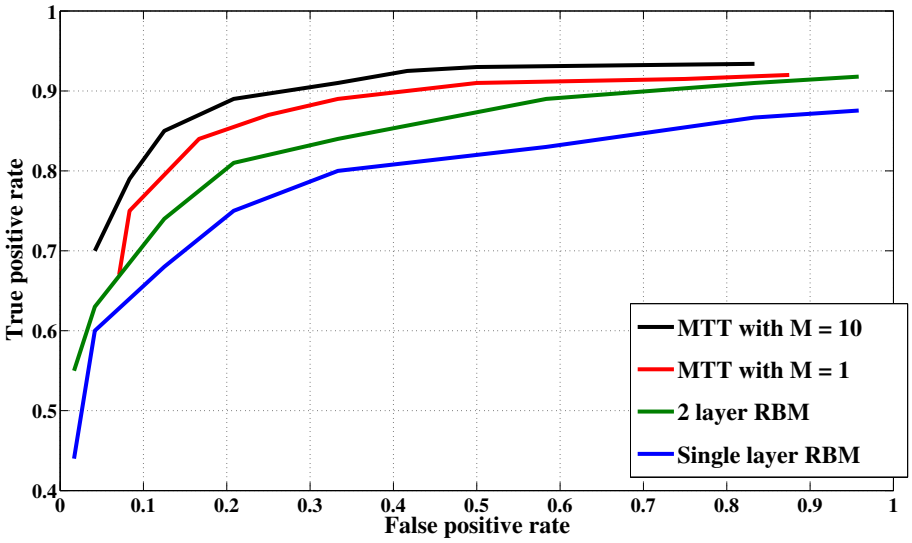

Figure 5.2: ROC curves for 4 systems regarding to detection of abnormal sound events

MTT confirms our intuition that MTT with desired number of components is able to segment an event with different modes of temporal trajectories into statistically salient sub-events.

We further test the robustness of the proposed system by adding noise from NOISEX-92 database [258] to the test set at different SNR levels of 20, 10, 0, -5 and -10 dB. The performance of the systems are measured in terms of percentage F-measure. We see from Table 5.1 that MTT (M=10) based detection system not only outperforms the other three system variants in clean scenario but exhibits robustness in presence of noise as well. When noise level increases, the detection performance of our MTT based system degrades at a much lower rate compared to the other three system variants. We also observe that for upto 10 dB SNR, our proposed system gives a very satisfactory performance in detecting the abnormal sound events. Another interesting point to

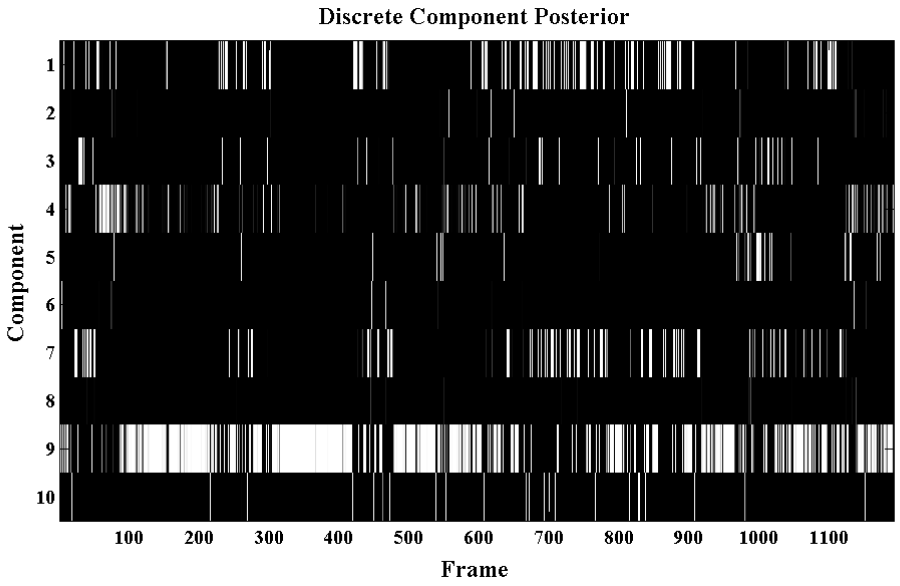

Figure 5.3: Posterior probability distribution of components corresponding to multiple "sub-events" in normal conversation

SNR (dB)	F-measure (%) for 4 detection systems			
	MTT (M=10)	MTT (M=1)	2 layer RBM	Single layer RBM
Clean	**93.11**	89.12	86.55	78.77
20 dB	**92.03**	85.41	79.66	71.82
10 dB	**88.85**	80.15	72.99	64.88
0 dB	**65.77**	59.87	51.66	43.77
-5 dB	**50.76**	43.28	34.88	25.75
-10 dB	**42.36**	34.88	25.77	10.99

Table 5.1: Abnormal sound events detection results for 4 systems at different SNR levels

note from Table 1 is that even MTT model with $M = 1$ performs better than RBM based models for all noise cases. This clearly shows the importance of incorporating the information of temporal trajectories along with localized spectro-temporal attributes in the model representation of sound events for a robust characterization.

5.4 Discussion

In this work, we develop a hybrid RBM-MTT framework for abnormal sound event detection in subway station by using a joint representation of localized spectro-temporal attributes with mixtures of temporal trajectories. Such a joint representation is very effective in capturing the intricate details and commonalities across a broader sound class spanned by multiple events. We show that MTT as a *dynamical* mixture model spans the complete temporal trajectory space of a complex acoustic scene by decomposing it into multiple trajectories, each of which belongs to a particular sub-event. In abnormal sound event detection task, the detection accuracy improves by an absolute 7 % over RBM class of models when information of different modes of temporal dynamics is incorporated in model representation of sound objects via our proposed MTT. We also find that our MTT based representation augments the detection system with high degree of noise robustness at low SNR levels, thus illustrating the fact that the joint representation provides a much robust characterization of broader sound classes.

Chapter 6

Conclusion

6.1 Overall conclusions

This thesis has explored the computational strategies underlying the hierarchical representation of sound in auditory pathway and its role in driving bottom-up and top-down mechanisms in driving several auditory scene analysis paradigms namely acoustic scene classification, salient event detection and scene segregation among others. The first part of the thesis explored the role of temporal dynamics of spectro-temporal modulation features in capturing intricate details in auditory scenes that extend beyond average statistics of the scene and track the heterogeneous dynamics commonly encountered in these scenes. Specifically, we proposed that temporal trajectories of local spectral temporal profiles do provide complimentary information in addition to their mean statistics. A fusion system based on both representations provides a better model of each sound class relative to the individual models able on mean statistics and temporal dynamics. We showed that such a hybrid systems is crucial in capturing non-trivial details from an unstructured and dynamic

acoustic environment. This was reflected in an acoustic scene classification paradigm in which the hybrid representation based on GMM-HMM framework showed significantly improved performance than baseline MFCC-GMM framework, hence validating the contribution of temporal dynamics information incorporated along with mean statistics of modulation features.

In the next part of the thesis, we have emphasized on demonstrating the ability of our model in learning a rich representation of sound from complex sound classes in an unsupervised fashion. The proposed computational framework replicates the physiological manifestation of auditory system with a multi-layered neural network architecture and closely resembles aspects of hierarchical transformations of sound in auditory pathway. The localized spectro-temporal bases representation and their long-term temporal regularities learned via a deep belief network capture the finer and global details in commensurate with the dynamics of the sound. These varying degree of details are then projected onto the space of an auditory object via a hebbian based grouping mechanism based in the theory of temporal coherence. We have demonstrated that such an integrative framework drive scene segregation processes in varied complexity of sounds ranting from simple tones to complex speech utterances. Another major advantage of this framework is its key relevance in terms of bridging the existing gap between physiological bases of how sound is represented in human brain and the psychoacoustic theories driving the processes of auditory scene analysis paradigm

The hierarchical acoustic representation in terms of localized and global spectro-temporal bases is then explored in the context of adaptation to natural scene statistics. We have demonstrated that such an adaptation framework

tunes the spectro-temporal bases as per change in temporal regularities of a scene. This framework forms the core of bottom-up saliency mechanism. We have demonstrated that the proposed bottom-up saliency framework achieves significantly improved performance over baseline feature driven bottom-up frameworks reported in previous studies for a abnormal event detection task. We have further supplemented the bottom-up framework with top-down knowledge of "events" of interest (abnormal in this case) from acoustic scene classification framework. The incorporation of top-down knowledge of abnormal events into bottom-up framework further improves the performance of the entire architecture in terms correct detection of events of interest. The top-down knowledge also imparts high degree of robustness to the architecture at low SNRs, thus illustrating the fact that bottom-up and top-down auditory factors provide a much robust characterization of complex abnormal events and capture the salient details of different degrees even in presence of highly prominent background events.

Finally, we extended the abnormal event detection paradigm with an aim to develop a global understanding of an unstructured and heterogeneous acoustic scene via a rich hierarchy based acoustic representation. We develop a hybrid RBM-MTT framework for abnormal sound event detection in subway station by using a joint representation of localized spectro-temporal attributes with mixtures of temporal trajectories. We show that such a joint representation is not only very effective in capturing the intricate details and commonalities across a broader sound class but also spans the complete temporal trajectory space of a complex acoustic scene by decomposing it into multiple trajectories corresponding to a specific semantic category within the acoustic scene.

6.2 Further Extensions of this Work

The computational architecture based on rich hierarchy of acoustic representation presented in this work demonstrates a reasonable performance for a number of scene analysis related tasks like acoustic scene classification, scene segregation and event detection among others. However, in a constantly changing acoustic environment, there needs to be a continuous two-way interaction between bottom-up and top-down stages in the hierarchy such that the bottom-up framework constantly adapts itself as per the prior knowledge of an acoustic environment.

As per the hierarchical architecture presented in Chapter 4, the integration of top-down knowledge into the bottom-up stage is primarily driven by acoustic scene classification framework. This architecture can be extended in a number of ways. First, both bottom-up and top-down components can be tuned together using a discriminative objective function, similar to adaptation frameworks used in speaker verification and speech recognition paradigms [259, 260]. The objective function may be designed in such a way that the bottom-up representation changes its characteristics based on heuristic knowledge of acoustic environment.

Second, it still remains to be seen whether induced plasticity in bottom-up representation using such a discriminative approach would correspond with well established neurophysiological results suggested in literature [261–263]. A number of psychoacoustic studies suggest that cognitive functions, such as attention and memory, drive perception by tuning sensory mechanisms to relevant acoustic features [264, 265]. An interesting extension of the architecture

proposed in this work will be to integrate it with cognitive mechanisms like attention and memory etc. to get an understanding of how such mechanisms drive scene analysis paradigms from computational viewpoint.

Finally, the results of Chapter 5 motivate further investigation into other practical scene analysis related paradigms. For example, it has been demonstrated in Chapter 5 that MTT framework is capable of decomposing an unstructured and heterogeneous scene into multiple trajectories, each of which belongs to specific sub-event forming a semantic category within the scene. Hence, it would be interesting to extend this framework to polyphonic event detection task in which multiple events with different trajectories are present at the same time instant. It is also of interest to explore the generative and discriminative MTT in greater depth to gather an understanding of whether the trajectories naturally emerge out from the acoustic representation of complex scene or the discriminative objective function forces the network to learn multiple trajectories corresponding to specific events within the scene.

Bibliography

[1] A S Bregman. *Auditory scene analysis: the perceptual organization of sound*. Cambridge, Mass.: MIT Press, 1990, pp. 1–773.

[2] Cynthia F. Moss and Annemarie Surlykke. "Auditory scene analysis by echolocation in bats". In: *The Journal of the Acoustical Society of America* 110.4 (2001), pp. 2207–2226.

[3] M. Elhilali, J.B. Fritz, D.J. Klein, J.Z. Simon, and S.A. Shamma. "Dynamics of Precise Spike Timing in Primary Auditory Cortex". In: *Journal of Neuroscience* 24.5 (2004), pp. 1159–1172.

[4] Jarmo Hurri and Aapo Hyvarinen. "Simple-cell-like receptive fields maximize temporal coherence in natural video". In: *Neural Comp.* 15 (2003), pp. 663–691.

[5] Li-jia Li, Hao Su, Eric P Xing, and Li Fei-fei. "Object Bank : A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification". In: *Nips* (2010), pp. 1–9.

[6] Liefeng Bo, X Ren, and Dieter Fox. "Kernel descriptors for visual recognition". In: *NIPS*. 2010, pp. 1–9.

[7] Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michael Mathieu, and Yann LeCun. "Learning Convolutional Feature Hierarchies for Visual Recognition". In: *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*. 2010.

[8] G.W. Taylor, L. Sigal, D.J. Fleet, and G.E. Hinton. "Dynamical binary latent variable models for 3D human pose tracking". In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), pp. 631–638.

[9] D Alais, R Blake, and S H Lee. "Visual features that vary together over time group together over space". In: *Nature neuroscience* 1.2 (1998), pp. 160–164.

[10] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Oliva. "Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks". In: *NeuroImage* 153 (2017), pp. 346–358. arXiv: arXiv:1011.1669v3.

[11] I Winkler, S L Denham, and I Nelken. "Modeling the auditory scene: predictive regularity representations and perceptual objects". In: *Trends in cognitive sciences* 13.12 (2009), p. 40.

[12] Alexandra Bendixen, Susan L Denham, Kinga Gyimesi, and István Winkler. "Regular patterns stabilize auditory streams". In: *Journal of the Acoustical Society of America* 128.6 (2010), pp. 3658–3666.

[13] I Nelken and O Bar-Yosef. "Neurons and objects: the case of auditory cortex". In: *Frontiers in neuroscience* 2.1 (2008), pp. 107–113.

[14] James O Pickles. *An Introduction to the Physiology of Hearing*. Third. Emerald Group Publishing Limited, 2008.

[15] J C Middlebrooks. "Auditory cortex cheers the overture and listens through the finale". In: *Nature neuroscience* 8.7 (2005), pp. 851–852.

[16] K J Friston. "Hierarchical models in the brain". In: *PLoS computational biology* 4.11 (2008), e1000211.

[17] Xiaoqin Wang. "The harmonic organization of auditory cortex". In: *Frontiers in Systems Neuroscience* 7 (2013).

[18] Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J. Oxenham, and Shihab A. Shamma. "Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes". In: *Neuron* 61.2 (2009), pp. 317–329.

[19] J T Geiger, B Schuller, and G Rigoll. "Large-scale audio feature extraction and SVM for acoustic scene classification". In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2013, pp. 1–4.

[20] Antti J. Eronen, Vesa T. Peltonen, Juha T. Tuomi, Anssi P. Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi. "Audio-based context recognition". In: *IEEE Transactions on Audio, Speech and Language Processing*. Vol. 14. 2006, pp. 321–329.

[21] Ieee Aasp Challenge and Acoustic Scenes. "A TONE-FIT FEATURE REPRESENTATION FOR SCENE CLASSIFICATION Johannes D . Krijnders , Gineke A . ten Holt 9401 HJ Assen The Netherlands". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 324.February 2016 (2013), pp. 359–367.

[22] Vesa Peltonen, Juha Tuomi, Anssi Klapuri, Jyri Huopaniemi, and Timo Sorsa. "Computational Auditory Scene Recognition". In: *IEEE International Conference on Audio, Speech and Signal Processing* (2002), pp. II–1941–II–1944.

[23] X Zhuang, X Zhou, T Huang, and Mark A Hasegawa-Johnson. "Feature Analysis and selection for acoustic event detection". In: *Proceedings of ICASSP'08*. 2008.

[24] Ozlem Kalinli, Shiva Sundaram, and Shrikanth 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.

[25] R J Zatorre and P Belin. "Spectral and temporal processing in human auditory cortex." In: *Cerebral cortex (New York, N.Y. : 1991)* 11.10 (2001), pp. 946–53.

[26] Howard Lei, Bernd T. Meyer, and Nikki Mirghafori. "Spectro-temporal Gabor features for speaker recognition". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4241–4244.

[27] M Carlin and M Elhilali. *An emergent spectro-temporal representation for natural sounds based on sustained firing of central auditory neurons*. Tech. rep. 2013.

[28] Courtenay V. Cotton and Daniel P W Ellis. "Spectral vs. spectro-temporal features for acoustic event detection". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2011, pp. 69–72.

[29] T Chi, P Ru, and S A Shamma. "Multiresolution spectrotemporal analysis of complex sounds". In: *The Journal of the Acoustical Society of America* 118.2 (2005), pp. 887–906.

[30] X Wang, T Lu, D Bendor, and E Bartlett. "Neural coding of temporal information in auditory thalamus and cortex". In: *Neuroscience* 157 (2008), pp. 484–493.

[31] Maneesh Sahani and Jennifer F Linden. “How linear are auditory cortical responses?” In: *Adv. Neural Inf. Proc. Sys. (NIPS)*. 2002.

[32] Kyogu Lee, Ziwon Hyung, and Juhan Nam. “Acoustic scene classification using sparse feature learning and event-based pooling”. In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[33] D Wang and P Chang. “An oscillatory correlation model of auditory streaming”. In: *Cognitive neurodynamics* 2.1 (2008), pp. 7–19.

[34] L P van Noorden and L P van Noorden. *Temporal coherence in the perception of tone sequences*. 1975.

[35] W Hartmann and D Johnson. “Stream segregation and peripheral channeling”. In: *Music Perception* 9.2 (1991), pp. 155–184.

[36] M W Beauvois and R Meddis. “A computer model of auditory stream segregation”. In: *Human experimental psychology* 43.3 (1991), pp. 517–541.

[37] S L McCabe and M J Denham. “A model of auditory streaming”. In: *Journal of the Acoustical Society of America* 101.3 (1997), pp. 1611–1621.

[38] Guoning Hu and DeLiang Wang. “A tandem algorithm for pitch estimation and voiced speech segregation”. In: *IEEE Transactions on Audio, Speech and Language Processing* 18.8 (2010), pp. 2067–2079.

[39] Lakshmi Krishnan, Mounya Elhilali, and Shihab Shamma. “Segregating complex sound sources through temporal coherence.” In: *PLoS computational biology* 10.12 (2014), e1003985.

[40] Chetan Singh Thakur, Runchun M. Wang, Saeed Afshar, Tara J. Hamilton, Jonathan C. Tapson, Shihab A. Shamma, and André van Schaik. “Sound stream segregation: A neuromorphic approach to solve the "cocktail party problem" in real-time”. In: *Frontiers in Neuroscience* 9.SEP (2015).

[41] C M Gray. “The temporal correlation hypothesis of visual feature integration: still alive and well.” In: *Neuron* 24.1 (1999), pp. 31–47, 111–25.

[42] Vesa T. K. Peltonen, Vesa T. K. Peltonen, Antti J. Eronen, Mikko P. Parviainen, and Anssi P. Klapuri. “Recognition of Everyday Auditory Scenes: Potentials, Latencies and Cues”. In: *IN PROC. 110TH AUDIO ENG. SOC. CONVENTION* (2001).

[43] Michael Buchler, Silvia Allegro, Stefan Launer, and Norbert Dillier. "Sound classification in hearing aids inspired by auditory scene analysis". In: *Eurasip Journal on Applied Signal Processing* 2005.18 (2005), pp. 2991–3002.

[44] Christophe Couvreur, Vincent Fontaine, Paul Gaunard, and Corine Ginette Mubikangiey. "Automatic classification of environmental noise events by hidden Markov models". In: *Applied Acoustics* 54.3 (1998), pp. 187–206.

[45] Shahrzad Esmaili, Sridhar Krishnan, and Kaamran Raahemifar. "Content based audio classification and retrieval using joint time-frequency analysis". In: *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. Vol. 5. IEEE, 2004, 8 vol. 5.

[46] E S Sussman. "Integration and segregation in auditory scene analysis". In: 117.3 (2005), pp. 1285–1298.

[47] Athanasia Zlatintsi, Elias Iosif, Petros Marago, and Alexandros Potamianos. "Audio salient event detection and summarization using audio and text modalities". In: *2015 23rd European Signal Processing Conference, EUSIPCO 2015*. 2015, pp. 2311–2315.

[48] Debmalya Chakrabarty and Mounya Elhilali. "Exploring the role of temporal dynamics in acoustic scene classification". In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[49] Gerard Roma, Waldo Nogueira, and Perfecto Herrera. "Recurrence quantification analysis features for environmental sound recognition". In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[50] Dimitrios Giannoulis, Dan Stowell, Emmanouil Benetos, Mathias Rossignol, Mathieu Lagrange, and Mark D. Plumbley. "A database and challenge for acoustic scene classification and event detection". In: *European Signal Processing Conference*. 2013. arXiv: arXiv:0706.3384v2.

[51] Xi Zhou, Xiaodan Zhuang, Ming Liu, Hao Tang, Mark Hasegawa-Johnson, and Thomas Huang. "HMM-Based Acoustic Event Detection with AdaBoost Feature Selection". In: *Multimodal Technologies for Perception of Humans*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 345–353.

[52] Chris J. Darwin. *Auditory grouping*. 1997.

[53] DeLiang DeLiang Wang. "Time-frequency masking for speech separation and its potential for hearing aid design." In: *Trends in amplification* 12.4 (2008), pp. 332–53.

[54] Tuomas Virtanen. "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria". In: *IEEE Transactions on Audio, Speech and Language Processing* 15.3 (2007), pp. 1066–1074.

[55] A. Cichocki, R. Zdunek, and S. Amari. "New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation". In: *IEEE Proceedings of the International Conference on Acoustics Speed and Signal Processing*. Vol. 5. 2006, pp. V–621–V–624. arXiv: 1010.1763.

[56] Guoxu Zhou, Andrzej Cichocki, Qibin Zhao, and Shengli Xie. "Non-negative matrix and tensor factorizations: An algorithmic perspective". In: *IEEE Signal Processing Magazine* 31.3 (2014), pp. 54–65. arXiv: arXiv:1011.1669v3.

[57] Rui Cai, Lie Lu, Alan Hanjalic, Hong Jiang Zhang, and Lian Hong Cai. "A flexible framework for key audio effects detection and auditory context inference". In: *IEEE Transactions on Audio, Speech and Language Processing* 14.3 (2006), pp. 1026–1038.

[58] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S. Lew. "Deep learning for visual understanding: A review". In: *Neurocomputing* 187 (2016), pp. 27–48.

[59] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. "Learning Deep Features for Scene Recognition using Places Database". In: *Advances in Neural Information Processing Systems 27* (2014), pp. 487–495. arXiv: 1504.05070.

[60] Clement Farabet, Camille Couprie, Laurent Najman, and Yann Lecun. "Learning hierarchical features for scene labeling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1915–1929. arXiv: arXiv:1011.1669v3.

[61] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition". In: *IEEE Transactions on Audio, Speech and Language Processing* 20.1 (2012), pp. 30–42. arXiv: 1201.0490.

[62] X. Zhao, Y. Wang, and D. Wang. "Deep neural networks for cochannel speaker identification". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2015-Augus. 2015.

[63] Yohan Petetin, Cyrille Laroche, and Aurelien Mayoue. "Deep neural networks for audio scene recognition". In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 125–129.

[64] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. "Polyphonic sound event detection using multi label deep neural networks". In: *2015 International Joint Conference on Neural Networks (IJCNN)*. 2015, pp. 1–7. arXiv: 1412.3555.

[65] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. "Multichannel audio source separation with deep neural networks". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 24.9 (2016), pp. 1652–1664.

[66] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. "Unsupervised feature learning for audio classification using convolutional deep belief networks". In: *Advances in neural information processing systems*. 2009, pp. 1096–1104.

[67] Aren Jansen, Jort F. Gemmeke, Daniel P.W. Ellis, Xiaofeng Liu, Wade Lawrence, and Dylan Freedman. "Large-scale audio event discovery in one million YouTube videos". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, 2017, pp. 786–790.

[68] Yonggang Qi, Jun Guo, Yi Li, Honggang Zhang, Tao Xiang, Yi Zhe Song, and Zheng Hua Tan. "Perceptual grouping via untangling Gestalt principles". In: *IEEE VCIP 2013 - 2013 IEEE International Conference on Visual Communications and Image Processing*. 2013.

[69] James K. Wright and Albert S. Bregman. "Auditory stream segregation and the control of dissonance in polyphonic music". In: *Contemporary Music Review* 2.1 (1987), pp. 63–92.

[70] Joel S Snyder and Claude Alain. "Toward a neurophysiological theory of auditory stream segregation". In: *Psychological bulletin* 133.5 (2007), pp. 780–799.

[71] Daniel P.W. Ellis and Keansub Lee. “Minimal-impact audio-based personal archives”. In: *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences - CARPE'04* (2004), p. 39.

[72] Michael Kleinschmidt. “Localized spectro-temporal features for automatic speech recognition”. In: *Proceedings of Eurospeech*. Citeseer, 2003.

[73] Waldo Nogueira, Gerard Roma, and Perfecto Herrera. “Automatic Event Classification Using Front End Single Channel Noise Reduction, MFCC Features and a Support Vector Machine Classifier”. In: *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*. 2013, pp. 1–2.

[74] J.D. Krijnders, M.E. Niessen, and T.C. Andringa. “Sound event recognition through expectancy-based evaluation ofsignal-driven hypotheses”. In: *Pattern Recognition Letters* 31.12 (2010), pp. 1552–1559.

[75] Kailash Patil and Mounya Elhilali. “Goal-Oriented Auditory Scene Recognition”. In: *13th Annual Conference of the International Speech Communication Association, INTERSPEECH 2012*. Vol. 3. 2012, pp. 2510–2513.

[76] Http://www.sound-ideas.com/bbc.html. *The BBC Sound Effects Library*. 1990.

[77] Marios Athineos, Hynek Hermansky, and Daniel P.W. Ellis. “PLP^2 fffdfffdfffd Autoregressive modeling of auditory-like 2-D spectro-temporal patterns”. In: *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing* (2004).

[78] Lee Tai Sing. “Image representation using 2D Gabor wavelets”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18.10 (1996), pp. 959–971.

[79] R L De Valois and Karen K De Valois. “Spatial Vision”. In: *Annu. Rev. Psychol.* 31 (1980), pp. 309–341.

[80] L De Lathauwer, B De Moor, J Vandewalle, L De Lathauwer, B De Moor, and J Vandewalle. “A multilinear singular value decomposition”. In: 21 (2000), pp. 1253–1278.

[81] Douglas A. Reynolds and Richard C. Rose. “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”. In: *IEEE Transactions on Speech and Audio Processing* 3.1 (1995), pp. 72–83.

[82] Kshitiz Kumar, Chanwoo Kim, and Richard M. Stern. “Delta-spectral cepstral coefficients for robust speech recognition”. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011, pp. 4784–4787.

[83] B. Clarkson and A. Pentland. “Extracting context from environmental audio”. In: *Digest of Papers. Second International Symposium on Wearable Computers*. IEEE Comput. Soc, pp. 154–155.

[84] L. Rabiner and B. Juang. “An introduction to hidden Markov models”. In: *IEEE ASSP Magazine* 3.January (1986), Appendix 3A.

[85] N Brummer, L Burget, J H Cernocky, O Glembek, F Grezl, M Karafiat, D A van Leeuwen, P Matejka, P Schwarz, and A Strasheim. “Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 15.7 (2007), pp. 2072–2084.

[86] Lewis O Harvey and Thomas D. Wickens. “Elementary Signal Detection Theory”. In: *Psychology of Perception* (2001), pp. 1–288. arXiv: arXiv:1011.1669v3.

[87] Edward W. Large and Caroline Palmer. “Perceiving temporal regularity in music”. In: *Cognitive Science* 26.1 (2002), pp. 1–37. arXiv: arXiv:1011.1669v3.

[88] T O Sharpee, C A Atencio, and C E Schreiner. “Hierarchical representations in the auditory cortex”. In: *Current opinion in neurobiology* 21.5 (2011), pp. 761–767.

[89] M S Lewicki, B A Olshausen, A Surlykke, and C F Moss. “Scene analysis in the natural environment”. In: *Frontiers in psychology* 5 (2014), p. 199.

[90] J M Wolfe, T S Horowitz, and N M Kenner. “Cognitive psychology: rare items often missed in visual searches”. In: 435.7041 (2005), pp. 439–440.

[91] C J Darwin. “Auditory grouping”. In: *Trends in cognitive sciences* 1.9 (1997), pp. 327–333.

[92] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der Heydt. “A century of Gestalt psychology in visual perception: I. Perceptual grouping and figurefffdfffdfffdground organization.” In: *Psychological Bulletin* 138.6 (2012), pp. 1172–1217. arXiv: NIHMS150003.

[93] S Haykin and Z Chen. “The cocktail party problem”. In: *Neural computation* 17.9 (2005), pp. 1875–1902.

[94] G. E. Hinton and R. R. Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *Science* 313.5786 (2006), pp. 504–507. arXiv: 20.

[95] T D Griffiths and J D Warren. “What is an auditory object?” In: *Nature neurosc.reviews* 5.11 (2004), pp. 887–892.

[96] D P W Ellis and R J Weiss. “Model-based monaural source separation using vector-quantized phase-vocoder representation”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 5. 2006, pp. 957–960.

[97] G J Jang and T W Lee. “A Maximum Likelihood Approach to Single-channel Source Separation”. In: *Journal of Machine Learning Research* 4.7-8 (2003), pp. 1365–1392.

[98] S Grossberg, K K Govindarajan, L L Wyse, and M A Cohen. “ARTSTREAM: a neural network model of auditory scene analysis and source segregation”. In: *Neural networks* 17.4 (2004), pp. 511–536.

[99] J Nix and V Hohmann. “Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering”. In: *IEEE Transactions on Audio, Speech and Language Processing* 15.3 (2007), pp. 995–1008.

[100] M Elhilali. “Modeling the cocktail party problem”. In: *The auditory system at the cocktail party*. Ed. by J Middlebrooks, J Simon, A Popper, and R Fay. New York, NY: Springer, 2017, pp. 111–135.

[101] Maximilian Riesenhuber and Tomaso Poggio. “Hierarchical models of object recognition in cortex”. In: *Nature neuroscience* 2.11 (1999), pp. 1019–1025.

[102] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Computation* 1.4 (1989), pp. 541–551.

[103] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-antoine Manzagol. “Deep Learning with Denoising Autoencoders”. In: *Journal of Machine Learning* 27 (2008), pp. 49–50.

[104] Huiyu Zhou, Yuan Yuan, and Chunmei Shi. “Object tracking using SIFT features and mean shift”. In: *Computer Vision and Image Understanding* 113.3 (2009), pp. 345–352.

[105] Benjamin W Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard. "Eye guidance in natural vision: Reinterpreting salience". In: *Journal of Vision* 11.5 (2011).

[106] Eric Nowak, Frédéric Jurie, and Bill Triggs. "Sampling strategies for bag-of-features image classification". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 3954 LNCS. 2006, pp. 490–503.

[107] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations (ICRL)* (2015), pp. 1–14. arXiv: 1409.1556.

[108] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1–8. arXiv: arXiv:1301.3605v3.

[109] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *arXiv* (2015), pp. 1–15. arXiv: 1511.06434.

[110] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, and Andrew Y. Ng. "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning". In: *2011 International Conference on Document Analysis and Recognition* (2011), pp. 440–445. arXiv: fa.

[111] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. "A Fast Learning Algorithm for Deep Belief Nets". In: *Neural Computation* 18.7 (2006), pp. 1527–1554. arXiv: 1111.6189v1.

[112] S A Shamma. "Spatial and temporal processing in central auditory networks". In: *Methods of Neuronal modeling: From ions to networks*. Ed. by C Koch and I Segev. second. MIT Press, Cambridge, 1998, pp. 411–460.

[113] S A Shamma and D J Klein. "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system". In: *Journal of the Acoustical Society of America* 107.5 (2000), pp. 2631–2644.

[114] X Yang, K Wang, and S A Shamma. "Auditory representations of acoustic signals". In: *IEEE Trans. Inf. Theory* 38.2 (1992), pp. 824–839.

[115] C E Schreiner. “Spatial distribution of responses to simple and complex sounds in the primary auditory cortex”. In: *Audiology and Neuro-otology* 3.2-3 (1998), pp. 104–122.

[116] DeLiang L. Wang and Guy J. Brown. “Separation of speech from interfering sounds based on oscillatory correlation”. In: *IEEE Transactions on Neural Networks* 10.3 (1999), pp. 684–697.

[117] Andrew J R Simpson, Gerard Roma, and Mark D. Plumbley. “Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9237. 2015, pp. 429–436. arXiv: 1504.04658.

[118] Debmalya Chakrabarty and Mounya Elhilali. “Abnormal sound event detection using temporal trajectories mixtures”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 216–220.

[119] Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip J.B. Jackson, and Mark D. Plumbley. “Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging”. In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 25.6 (2017), pp. 1230–1241. arXiv: 1607.03681.

[120] Tara N. Sainath, Dimitri Kanevsky, and Giridharan Iyengar. “Unsupervised audio segmentation using extended Baum-Welch transformations”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 1. 2007, pp. 209–212.

[121] C J Darwin and R P Carlyon. “Auditory grouping”. In: *Trends in cognitive sciences*. Ed. by B C J Moore. Vol. 6. Hearing. Orlando, FL: Academic Press, 1995, pp. 387–424.

[122] Barbara Shinn-Cunningham, Adrian K C Lee, and Andrew J Oxenham. “A sound element gets lost in perceptual competition”. In: *Proc.Nat.Acad.Sci.* 104.29 (2007), pp. 12223–12227.

[123] Chris J Darwin. “Simultaneous Grouping and Auditory Continuity”. In: *Perception & Psychoacoustics* 67.8 (2005), pp. 1384–1390.

[124] C J Darwin, R W Hukin, and B Y Al-Khatib. “Grouping in pitch perception: Evidence for sequential constraints”. In: *Journal of the Acoustical Society of America* 98.2 I (1995), pp. 880–885.

[125] B C J Moore and H Gockel. “Factors influencing sequential stream segregation”. In: *Acta Acustica* 88 (2002), pp. 320–333.

[126] Titia L. van Zuijen, Elyse Sussman, István Winkler, Risto Näätänen, and Mari Tervaniemi. “Grouping of Sequential SoundsfffdfffdfffdAn Event-Related Potential Study Comparing Musicians and Nonmusicians”. In: *Journal of Cognitive Neuroscience* 16.2 (2004), pp. 331–338.

[127] Virginia Best, Frederick J. Gallun, Simon Carlile, and Barbara G. Shinn-Cunningham. “Binaural interference and auditory grouping”. In: *The Journal of the Acoustical Society of America* 121.2 (2007), pp. 1070–1076.

[128] Kamil Hamaoui and Diana Deutsch. “The perceptual grouping of musical Sequences: Pitch and timing as competing cues”. In: *Proceedings of the 11th International Conference on Music Perception and Cognition* 11 (2011), pp. 81–87.

[129] Matthew Luciw and Juyang Weng. “Top-down connections in self-organizing hebbian networks: Topographic class grouping”. In: *IEEE Transactions on Autonomous Mental Development* 2.3 (2010), pp. 248–261.

[130] Michael S Falconbridge, Robert L Stamps, and David R Badcock. “A simple Hebbian/anti-Hebbian network learns the sparse, independent components of natural images.” In: *Neural computation* 18.2 (2006), pp. 415–29.

[131] Xiaohui Xie and H. Sebastian Seung. “Equivalence of Backpropagation and Contrastive Hebbian Learning in a Layered Network”. In: *Neural Computation* 15.2 (2003), pp. 441–454.

[132] R R Fay. “Perception of spectrally and temporally complex sounds by the goldfish (Carassius auratus)”. In: *Hearing research* 89.1-2 (1995), pp. 146–154.

[133] Justin Salamon and Juan Pablo Bello. “Unsupervised feature learning for urban sound classification”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2015-Augus. 2015, pp. 171–175. arXiv: arXiv:1011.1669v3.

[134] Geoffrey E. Hinton. “Training Products of Experts by Minimizing Contrastive Divergence”. In: *Neural Computation* 14.8 (2002), pp. 1771–1800.

[135] M Elhilali, S A Shamma, J Z Simon, and J B Fritz. “A Linear Systems View to the Concept of STRF”. In: *Handbook of Modern Techniques in Auditory Cortex*. Ed. by D Depireux and M Elhilali. Nova Science Pub Inc, 2013, pp. 33–60.

[136] Graham W. Taylor and Geoffrey E. Hinton. “Factored conditional restricted Boltzmann Machines for modeling motion style”. In: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. 2009, pp. 1–8.

[137] Shihab A. Shamma, Mounya Elhilali, and Christophe Micheyl. “Temporal coherence and attention in auditory scene analysis”. In: *Trends in neurosciences* 34.3 (2011), pp. 114–23.

[138] Shihab Shamma, Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J. Oxenham, Daniel Pressnitzer, Pingbo Yin, and Yanbo Xu. “Temporal coherence and the streaming of complex sounds”. In: *Advances in experimental medicine and biology* 787 (2013), pp. 535–43.

[139] J. J. Hopfield. “Neural networks and physical systems with emergent collective computational abilities.” In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558.

[140] A. J. Storkey and R. Valabregue. “The basins of attraction of a new Hopfield learning rule”. In: *Neural Networks* 12.6 (1999), pp. 869–876.

[141] N Singh and F Theunissen. “Modulation spectra of natural sounds and ethological theories of auditory processing”. In: *J.Acoust.Soc.Am.* 106 (2003), pp. 3394–3411.

[142] T M Elliott and F E Theunissen. “The modulation transfer function for speech intelligibility”. In: *PLoS computational biology* 5.3 (2009), e1000302.

[143] S A Shamma, H Versnel, and N Kowalski. “Ripple Analysis in Ferret Primary Auditory Cortex. I. Response Characteristics of Single Units to Sinusoidally Rippled Spectra”. In: *Institute for Systems Research Technical Reports* (1994).

[144] C Schreiner and B Calhoun. “Spectral envelope coding in cat primary auditory cortex: Properties of ripple transfer functions”. In: *J.Aud.Neurosc.* 1 (1995), pp. 39–61.

[145] M Schonwiesner and R J Zatorre. “Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.34 (2009), pp. 14611–14616.

[146] D A Depireux, J Z Simon, D J Klein, and S A Shamma. “Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex”. In: *Journal of neurophysiology* 85.3 (2001), pp. 1220–1234.

[147] Lee M. Miller, Monty A. Escabí, Heather L. Read, Christoph E. Schreiner, M A Escabi, Heather L. Read, and Christoph E. Schreiner. “Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex”. In: *Journal of neurophysiology* 87.1 (2002), pp. 516–527.

[148] Monty A Escabi, Lee M Miller, Heather L Read, and Christoph E Schreiner. “Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus.” In: *Journal of Neuroscience* 23.37 (2003), pp. 11489–11504.

[149] Babajide O. Ayinde and Jacek M. Zurada. “Clustering of receptive fields in Autoencoders”. In: *Proceedings of the International Joint Conference on Neural Networks*. Vol. 2016-Octob. 2016, pp. 1310–1317.

[150] A S Bregman and A I Rudnicky. “Auditory segregation: stream or streams?” In: *Journal of Experimental Psychology-Human Perception and Performance* 1.3 (1975), pp. 263–267.

[151] L P van Noorden and L P van Noorden. “Minimun differences of level and frequency for perceptual fission of tone sequences ABAB”. In: *The Journal of the Acoustical Society of America* 61.4 (1977), pp. 1041–1045.

[152] C Micheyl, C Hanson, L Demany, S Shamma, and A J Oxenham. “Auditory stream segregation for alternating and synchronous tones”. In: *Journal of experimental psychology.Human perception and performance* 39.6 (2013), pp. 1568–1580.

[153] David Marvin Green and John A Swets. *Signal detection theory and psychophysics*. Vol. 1. Wiley New York, 1966.

[154] N A Macmillan and M Schwartz. “A probe-signal investigation of uncertain-frequency detection”. In: *J Acoust Soc Am* 58.5 (1975), pp. 1051–1058.

[155] R Naatanen, M Tervaniemi, E Sussman, P Paavilainen, and I Winkler. “"Primitive intelligence" in the auditory cortex”. In: *Trends in neurosciences* 24.5 (2001), pp. 283–288.

[156] C Micheyl, B Tian, R P Carlyon, and J P Rauschecker. “Perceptual organization of tone sequences in the auditory cortex of awake macaques”. In: *Neuron* 48.1 (2005), pp. 139–148.

[157] R P Carlyon. “How the brain separates sounds”. In: *Trends in cognitive sciences* 8.10 (2004), pp. 465–471.

[158] V Ciocca. “The auditory organization of complex sounds”. In: *Frontiers in bioscience : a journal and virtual library* 13 (2008), pp. 148–169.

[159] N Grimault, S P Bacon, and C Micheyl. “Auditory stream segregation on the basis of amplitude-modulation rate”. In: *The Journal of the Acoustical Society of America* 111.3 (2002), pp. 1340–1348.

[160] C Micheyl, H Kreft, S Shamma, and A J Oxenham. “Temporal coherence versus harmonicity in auditory stream formation”. In: *Journal of the Acoustical Society of America* 133.3 (2013), EL188–EL194.

[161] Robert S Bolia, W Todd Nelson, Mark A Ericson, and Brian D Simpson. “A Speech Corpus for Multitalker Communications Research”. In: *The Journal of the Acoustical Society of America* 107.2 (2000), pp. 1065–1066.

[162] Douglas S. Brungart. “Evaluation of speech intelligibility with the coordinate response measure”. In: *The Journal of the Acoustical Society of America* 109.5 (2001), pp. 2276–2279.

[163] David A. Eddins and Chang Liu. “Psychometric properties of the coordinate response measure corpus with various types of background interference”. In: *The Journal of the Acoustical Society of America* 131.2 (2012), EL177–EL183.

[164] J O Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, 1988.

[165] Ben Scholl, Xiang Gao, and Michael Wehr. “Nonoverlapping Sets of Synapses Drive On Responses and Off Responses in Auditory Cortex”. In: *Neuron* 65.3 (2010), pp. 412–421. arXiv: NIHMS150003.

[166] I Nelken, J K Bizley, F R Nodal, B Ahmed, A J King, and J W Schnupp. “Responses of auditory cortex to complex stimuli: functional organization revealed using intrinsic optical signals”. In: *Journal of neurophysiology* 99.4 (2008), pp. 1928–1941.

[167] Monty A. Escabi and Christoph E. Schreiner. “Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain.” In: *Journal of Neuroscience* 22.10 (2002), pp. 4114–31.

[168] S M N Woolley, Thane E Fremouw, Anne Hsu, and Frederic E Theunissen. “Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds”. In: *Nature Neurosci.* 8.10 (2005), pp. 1371–1379.

[169] H Liang, S L Bressler, M Ding, R Desimone, and P Fries. "Temporal dynamics of attention-modulated neuronal synchronization in macaque V4". In: *Neurocomputing* 52-54 (2003), p. 481.

[170] M Zeitler, P Fries, and S Gielen. "Assessing neuronal coherence with single-unit, multi-unit, and local field potentials". In: *Neural Comp.* 18.9 (2006), pp. 2256–2281.

[171] S Teki, N Barascud, S Picard, C Payne, T D Griffiths, and M Chait. "Neural Correlates of Auditory Figure-Ground Segregation Based on Temporal Coherence". In: *Cerebral cortex* 26.9 (2016), pp. 3669–3680.

[172] Kai Lu, Yanbo Xu, Pingbo Yin, Andrew J. Oxenham, Jonathan B. Fritz, and Shihab A. Shamma. "Temporal coherence structure rapidly shapes neuronal interactions". In: *Nature communications* 8 (2017), p. 13900.

[173] J L Goldstein. "An optimum processor theory for the central formation of the pitch of complex tones". In: *Journal of the Acoustical Society of America* 54 (1973), pp. 1496–1516.

[174] A. J. Oxenham, J. G. W. Bernstein, and H. Penagos. "Correct tonotopic representation is necessary for complex pitch perception". In: *Proceedings of the National Academy of Sciences* 101.5 (2004), pp. 1421–1425.

[175] Andrei S Kozlov and Timothy Q Gentner. "Central auditory neurons have composite receptive fields". In: *Proceedings of the National Academy of Sciences* 113.5 (2016), pp. 1441–1446.

[176] J C Middlebrooks, R W Dykes, and M M Merzenich. "Binaural response-specific bands in primary auditory cortex (AI) of the cat: topographical organization orthogonal to isofrequency contours". In: *Brain research* 181.1 (1980), pp. 31–48.

[177] Mounya Elhilali and Shihab A Shamma. "A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation". In: *The Journal of the Acoustical Society of America* 124.6 (2008), pp. 3751–71.

[178] David J Klein, Peter Konig, and Konrad P Kording. "Sparse spectrotemporal coding of sounds". In: *EURASIP J.Appl.Sig.Proc.* 2003.7 (2003), pp. 659–667.

[179] Nicole L Carlson and Michael Robert DeWeese Vivienne L. Ming. "Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus". In: *PLoS Comp.Bio.* 8.7 (2012), e1002594.

[180] H Hermansky and N Morgan. "RASTA Processing of Speech". In: *IEEE Trans.Speech and Audio Process.* 2.4 (1994), pp. 382–395.

[181] Sridhar Krishna Nemala, Kailash Patil, and Mounya Elhilali. "A Multistream Feature Framework Based on Bandpass Modulation Filtering for Robust Speech Recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.2 (2013), pp. 416–426.

[182] David (Motorola Labs Uk) Pearse, Hans-günter (Ericson Eurolab Deutschland GmbH) Hirsch, D Pearce, H. Hirsch, and D Pearce. "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions". In: *ICSLP 2000 (6th International Conference on Spokel Language Processing)*. Vol. 6. 2000, pp. 16–19.

[183] R Blake and S H Lee. "The role of temporal structure in human vision". In: *Behavioral and cognitive neuroscience review* 4.1 (2005), pp. 21–42.

[184] M W Beauvois and R Meddis. "Computer simulation of auditory stream segregation in alternating-tone sequences". In: *The Journal of the Acoustical Society of America* 99.4 (1996), pp. 2270–2280.

[185] M A Bee and G M Klump. "Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain". In: *Journal of neurophysiology* 92.2 (2004), pp. 1088–1104.

[186] D Pressnitzer, M Sayles, C Micheyl, and I M Winter. "Perceptual organization of sound begins in the auditory periphery". In: *Current Biology* 18.15 (2008), pp. 1124–1128.

[187] C Micheyl, R P Carlyon, A Gutschalk, J R Melcher, A J Oxenham, J P Rauschecker, B Tian, and E Courtenay Wilson. "The role of auditory cortex in the formation of auditory streams". In: *Hearing Research* 229.1-2 (2007), pp. 116–131.

[188] M Elhilali, L Ma, C Micheyl, A J Oxenham, and S A Shamma. "Rate vs. temporal code? A spatio-temporal coherence model of the cortical basis of streaming". In: *The Neurophysiological Bases of Auditory Perception*. Ed. by E Lopez-Poveda, A Palmer, and R Meddis. New York: Springer, 2010, pp. 497–506.

[189] V Ciocca and C J Darwin. "Effects of onset asynchrony on pitch perception: adaptation or grouping?" In: *Journal of the Acoustical Society of America* 93.5 (1993), pp. 2870–2878.

[190] Brian Roberts, Brian R Glasberg, and Brian C J Moore. "Effects of the build-up and resetting of auditory stream segregation on temporal discrimination". In: *Journal of Experimental Psychology: Human Perception and Performance* 34.4 (2007), pp. 992–1006.

[191] N R Haywood and B Roberts. "Build-up of the tendency to segregate auditory streams: Resetting effects evoked by a single deviant tone". In: *Journal of the Acoustical Society of America* 128.5 (2010), pp. 3019–3031.

[192] Susann Deike, Peter Heil, Martin Böckmann-Barthel, and André Brechmann. "The build-up of auditory stream segregation: a different perspective". In: *Frontiers in Psychology* 3 (2012), pp. 1–7.

[193] D Bendor and X Wang. "The neuronal representation of pitch in primate auditory cortex". In: *Nature* 436.7054 (2005), pp. 1161–1165.

[194] A Gutschalk, A J Oxenham, C Micheyl, E C Wilson, and J R Melcher. "Human cortical activity during streaming without spectral cues suggests a general neural substrate for auditory stream segregation". In: *Journal of Neuroscience* 27.48 (2007), pp. 13074–13081.

[195] C E Schreiner and M L Sutter. "Topography of excitatory bandwidth in cat primary auditory cortex: single-neuron versus multiple-neuron recordings". In: *Journal of neurophysiology* 68.5 (1992), pp. 1487–1502.

[196] D Bendor and X Wang. "Differential neural coding of acoustic flutter within primate auditory cortex". In: *Nature neuroscience* 10.6 (2007), pp. 763–771.

[197] Nicholas A. Smith and Suyash Joshi. "Neural correlates of auditory stream segregation: An analysis of onset- and change-related responses". In: *The Journal of the Acoustical Society of America* 136.4 (2014), EL295–EL301.

[198] R C DeCharms, D T Blake, and M M Merzenich. "Optimizing sound features for cortical neurons". In: *Science (New York, N.Y.)* 280.5368 (1998), pp. 1439–1443.

[199] Anthony Bell and Terrence Sejnowski. "Learning the higher-order structure of a natural sound". In: *Network: Computation in Neural Systems* 7.2 (1996), pp. 261–266.

[200] Claude Alain and Lori J. Bernstein. *From sounds to meaning: The role of attention during auditory scene analysis*. 2008.

[201] L Itti, C Koch, and E Niebur. “A model of saliency-based visual attention for rapid scene analysis”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 20.11 (1998), pp. 1254–1259.

[202] Charles E. Connor, Howard E. Egeth, and Steven Yantis. *Visual attention: Bottom-up versus top-down*. 2004.

[203] Hans Christoph Nothdurft. “Salience of feature contrast”. In: *Neurobiology of Attention*. 2005, pp. 233–239.

[204] C M Masciocchi, S Mihalas, D Parkhurst, and E Niebur. “Everyone knows what is interesting: Salient locations which should be fixated”. In: *Journal of Vision* 9.11 (2009), pp. 1–22.

[205] L Itti and C Koch. “Computational modelling of visual attention.” In: *Nature reviews. Neuroscience* 2.3 (2001), pp. 194–203. arXiv: arXiv:1011.1669v3.

[206] H J Seo and P Milanfar. “Static and space-time visual saliency detection by self-resemblance”. In: *Journal of Vision* 9.12 (2009), p. 15.

[207] X Hou and L Zhang. “Saliency detection: A spectral residual approach”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2007).

[208] N Bruce and J Tsotsos. “Saliency Based on Information Maximization”. In: *Advances in Neural Information Processing Systems 18* (2005), pp. 155–162.

[209] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. “SUN: A Bayesian framework for saliency using natural statistics”. In: *Journal of Vision* 8.7 (2008), p. 32.

[210] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. “Top-down control of visual attention in object detection”. In: *IEEE International Conference on Image Processing, September 14-17* 1 (2003), pp. 1–4. arXiv: arXiv:1011.1669v3.

[211] A Borji and L Itti. “State-of-the-art in visual attention modeling”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 185–207.

[212] Ali Borji, Majid Nili Ahmadabadi, Babak Nadjar Araabi, and Mandana Hamidi. “Online learning of task-driven object-based visual attention control”. In: *Image and Vision Computing* 28.7 (2010), pp. 1130–1145.

[213] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. 2012, pp. 1097–1105.

[214] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 07-12-June. 2015, pp. 1–9. arXiv: 1409.4842.

[215] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778. arXiv: 1512.03385.

[216] Xun Shi, Neil D.B. Bruce, and John K. Tsotsos. "Fast, recurrent, attentional modulation improves saliency representation and scene recognition". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2011.

[217] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. "Saliency Detection with Recurrent Fully Convolutional Networks". In: *Eccv*. 2016, pp. 825–841. arXiv: 1605.08110.

[218] Guanbin Li and Yizhou Yu. "Visual saliency detection based on multiscale deep CNN features". In: *IEEE Transactions on Image Processing* 25.11 (2016), pp. 5012–5024. arXiv: 1609.02077.

[219] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. "Saliency detection by multi-context deep learning". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 07-12-June. 2015, pp. 1265–1274.

[220] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis. "Video event detection and summarization using audio, visual and text saliency". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2009, pp. 3553–3556.

[221] José Portêlo, Miguel Bugalho, Isabel Trancoso, João Neto, Alberto Abad, and António Serralheiro. "Non-speech audio event detection". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2009, pp. 1973–1976.

[222] S Nemala and M Elhilali. *Relevant spectro-temporal modulations for robust speech and nonspeech classification*. Tech. rep. 2010.

[223] Hardik B. Sailor, Dharmesh M. Agrawal, and Hemant A. Patil. "Unsupervised filterbank learning using Convolutional Restricted Boltzmann Machine for environmental sound classification". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2017-Augus. 2017, pp. 3107–3111.

[224] Hamid Eghbal-Zadeh. "Cp-jku submissions for dcase-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks". In: *Detection and Classification of Acoustic Scenes and Events - DCASE2016 IEEE AASP Challenge* September (2016).

[225] Soo Hyun Bae, Inkyu Choi, and Nam Soo Kim. "Acoustic Scene Classification Using Parallel Combination of LSTM and CNN". In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)* September (2016), pp. 11–15.

[226] R D Melara and L E Marks. "Perceptual primacy of dimensions: support for a model of dimensional interaction". In: *Journal of Experimental Psychology: Human Perception and Performance* 16.2 (1990), pp. 398–414.

[227] E J Allen and A J Oxenham. "Interactions of Pitch and Timbre: How Changes in One Dimension Affect Discrimination of the Other". In: *Abstracts of the 36th ARO Mid-Winter meeting: Association of Research Otolaryngologists*. Vol. 36. 2013.

[228] Dan Wang and Yi Shang. "Modeling Physiological Data with Deep Belief Networks." In: *International journal of information and education technology (IJIET)* 3.5 (2013), pp. 505–511. arXiv: NIHMS150003.

[229] Emine Merve Kaya and Mounya Elhilali. "Investigating bottom-up auditory attention". In: *Frontiers in Human Neuroscience* 8 (2014), p. 327.

[230] Konrad P. Kording, Joshua B. Tenenbaum, and Reza Shadmehr. "The dynamics of memory as a consequence of optimal adaptation to a changing body". In: *Nature Neuroscience* 10.6 (2007), pp. 779–786.

[231] Geoffrey E. Hinton. "A Practical Guide to Training Restricted Boltzmann Machines". In: *Neural Networks: Tricks of the Trade*. Vol. 7700. 2012, pp. 599–619.

[232] Xin Li, Feipeng Zhao, and Yuhong Guo. "Conditional Restricted Boltzmann Machines for Multi-label Learning with Incomplete Labels". In: *Aistats* 38 (2015), pp. 635–643.

[233] E.A. Wan and R Van Der Merwe. "The unscented Kalman filter for nonlinear estimation". In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*. 2002, pp. 153–158.

[234] Gregory Gancarz and Stephen Grossberg. *A neural model of saccadic eye movement control explains task-specific adaptation*. 1999.

[235] Alex Graves and Jürgen Schmidhuber. "Framewise phoneme classification with bidirectional LSTM networks". In: *Proceedings of the International Joint Conference on Neural Networks*. Vol. 4. 2005, pp. 2047–2052.

[236] J S Garofolo, L F Lamel, W M Fisher, J G Fiscus, D S Pallett, N L Dahlgren, and V Zue. *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus*. Tech. rep. 1993.

[237] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. "Reliable detection of audio events in highly noisy environments". In: *Pattern Recognition Letters* 65 (2015), pp. 22–28.

[238] Karol J. Piczak. "ESC: Dataset for Environmental Sound Classification". In: *Proceedings of the 23rd ACM International Conference on Multimedia, MM 2015* (2015), pp. 1015–1018.

[239] Matthew D. Zeiler. "KIT08-ADADELTA: An Adaptive Learning Rate Method". In: *arXiv* (2012), p. 6. arXiv: 1212.5701.

[240] Nicholas Huang and Mounya Elhilali. "Auditory salience using natural soundscapes". In: *The Journal of the Acoustical Society of America* 141.3 (2017), p. 2163.

[241] Karol J. Piczak. "Environmental sound classification with convolutional neural networks". In: *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*. Vol. 2015-Novem. 2015.

[242] Yuji Tokozume and Tatsuya Harada. "Learning environmental sounds with end-to-end convolutional neural network". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2017, pp. 2721–2725.

[243] Kyungtae Kim, Kai-Hsiang Lin, Dirk B Walther, Mark A Hasegawa-Johnson, and Tomas S Huang. "Automatic detection of auditory salience with optimized linear filters derived from human annotation". In: *Pattern Recognition Letters* 38 (2014), pp. 78–85.

[244] V Duangudom and D V Anderson. "Using Auditory Saliency To Understand Complex Auditory Scenes". In: *15th European Signal Processing Conference (EUSIPCO 2007)*. 2007.

[245] T Tsuchida and G Cottrell. *Auditory saliency using natural statistics*. 2012.

[246] O Kalinli and S Narayanan. "A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech". In: *INTERSPEECH-2007*. 2007, pp. 1941–1944.

[247] Vinay Kumar Mittal and B. Yegnanarayana. "Production features for detection of shouted speech". In: *2013 IEEE 10th Consumer Communications and Networking Conference, CCNC 2013*. 2013, pp. 106–111.

[248] Mahesh Kumar Nandwana, Ali Ziaei, and John H L Hansen. "Robust unsupervised detection of human screams in noisy acoustic environments". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2015-Augus. 2015, pp. 161–165.

[249] Mary Knox and Nikki Mirghafori. "Automatic laughter detection using neural networks". In: *Proceedings of INTERSPEECH'07* (2007).

[250] S Ntalampiras, I Potamitis, and N Fakotakis. "On Acoustic Surveillance of Hazardous Situations". In: *Proceedings of ICASSP 2009* (2009), pp. 165–168.

[251] Geoffrey E. Hinton. *Learning multiple layers of representation*. 2007. arXiv: 1102.0183.

[252] Weimin Huang, Tuan Kiang Chiew, Haizhou Li, Tian Shiang Kok, and Jit Biswas. "Scream detection for home applications". In: *Proceedings of the 2010 5th IEEE Conference on Industrial Electronics and Applications, ICIEA 2010*. 2010, pp. 2115–2120.

[253] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. "Scream and gunshot detection and localization for audio-surveillance systems". In: *2007 IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007 Proceedings*. 2007, pp. 21–26.

[254] Graham W Taylor, Geoffrey E Hinton, and Sam Roweis. "Modeling Human Motion Using Binary Latent Variables". In: *Advances in Neural Information Processing Systems (NIPS)* (2007), pp. 1345–1352.

[255] Marc Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann Lecun. "Efficient Learning of Sparse Representations with an Energy-Based Model". In: *Advances In Neural Information Processing Systems* 19 (2007), pp. 1137–1134.

[256] KyungHyun Cho, Tapani Raiko, and Alexander Ilin. "Enhanced Gradient for Training Restricted Boltzmann Machines". In: *Neural Computation* 25.3 (2013), pp. 805–831.

[257] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila. "CASSANDRA: Audio-video sensor fusion for aggression detection". In: *2007 IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007 Proceedings*. 2007, pp. 200–205.

[258] A P Varga, H J M Steeneken, M Tomlinson, and D Jones. "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition". In: *Tech.Rep., Speech Research Unit, Defense Research Agency, Malvern, U.K.* (1992).

[259] M J F Gales and P C Woodland. "Mean and variance adaptation within the MLLR framework". In: *Computer Speech & Language* 10.4 (1996), pp. 249–264.

[260] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide. "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7893–7897.

[261] Jonathan Fritz, Shihab Shamma, Mounya Elhilali, and David Klein. "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex". In: *Nature Neuroscience* 6.11 (2003), pp. 1216–1223.

[262] Peter Lakatos, Gabriella Musacchia, Monica N. O'Connel, Arnaud Y. Falchier, Daniel C. Javitt, and Charles E. Schroeder. "The Spectrotemporal Filter Mechanism of Auditory Selective Attention". In: *Neuron* 77.4 (2013), pp. 750–761.

[263] Mounya Elhilali, Juanjuan Xiang, Shihab A. Shamma, and Jonathan Z. Simon. "Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene". In: *PLoS Biology* 7.6 (2009). Ed. by Timothy D. Griffiths, e1000129.

[264] Joel S. Snyder and Melissa K. Gregg. "Memory for sound, with an ear toward hearing in complex auditory scenes". In: *Attention, Perception, & Psychophysics* 73.7 (2011), pp. 1993–2007.

[265] Adelbert W. Bronkhorst. *The cocktail-party problem revisited: early processing and selection of multi-talker speech*. 2015.

Vita

Debmalya Chakrabarty (b. 1988) received the B.Tech degree in Electrical and Communications Engineering from National Institute of Technology Silchar, India in 2006 and M.S.E degree in Electrical and Computer Engineering from Johns Hopkins University in 2014. Prior to enrolling in the Ph.D. program in Electrical and Computer Engineering at Johns Hopkins, he worked as Configuration Engineer at Ericsson Global Service Ltd. in India. He also worked as a Research assistant in the department of Electrical Engineering at IIT Guwahati under the supervision of Dr. S.R.M. Prasanna. While a graduate student, he worked as a research intern at Xerox Research Center India in 2014. His research focuses on developing computational strategies based on hierarchical representation of sound in auditory pathway, and how to apply such strategies for solving various auditory scene analysis related paradigms.
Starting in June 2018, Debmalya will work as a Research Scientist with Amazon in Bangalore, India