

## ABSTRACT

Title of dissertation: NEURAL BASIS AND COMPUTATIONAL STRATEGIES  
FOR AUDITORY PROCESSING

Mounya Elhilali, Doctor of Philosophy, 2004

Dissertation directed by: Professor Shihab A. Shamma  
Department of Electrical and Computer Engineering

Our senses are our window to the world, and hearing is the window through which we perceive the world of sound. While seemingly effortless, the process of hearing involves complex transformations by which the auditory system consolidates acoustic information from the environment into perceptual and cognitive experiences. Studies of auditory processing try to elucidate the mechanisms underlying the function of the auditory system, and infer computational strategies that are valuable both clinically and intellectually, hence contributing to our understanding of the function of the brain.

In this thesis, we adopt both an experimental and computational approach in tackling various aspects of auditory processing. We first investigate the neural basis underlying the function of the auditory cortex, and explore the dynamics and computational mechanisms of cortical processing. Our findings offer physiological evidence for a role of primary cortical neurons in the integration of sound features at different time constants, and possibly in the formation of auditory objects.

Based on physiological principles of sound processing, we explore computational implementations in tackling specific perceptual questions. We exploit our knowledge of the neural mechanisms of cortical auditory processing to formulate models addressing the problems of speech intelligibility and auditory scene analysis. The intelligibility model

focuses on a computational approach for evaluating loss of intelligibility, inspired from mammalian physiology and human perception. It is based on a multi-resolution filter-bank implementation of cortical response patterns, which extends into a robust metric for assessing loss of intelligibility in communication channels and speech recordings.

This same cortical representation is extended further to develop a computational scheme for auditory scene analysis. The model maps perceptual principles of auditory grouping and stream formation into a computational system that combines aspects of bottom-up, primitive sound processing with an internal representation of the world. It is based on a framework of unsupervised adaptive learning with Kalman estimation. The model is extremely valuable in exploring various aspects of sound organization in the brain, allowing us to gain interesting insight into the neural basis of auditory scene analysis, as well as practical implementations for sound separation in “cocktail-party” situations.

NEURAL BASIS AND COMPUTATIONAL STRATEGIES  
FOR AUDITORY PROCESSING

by

Mounya Elhilali

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2004

Advisory Committee:

Professor Shihab A. Shamma, Chairman/Advisor  
Professor K. J. Ray Liu  
Professor James A. Reggia  
Assistant Professor Jonathan Z. Simon  
Research Scientist Christophe Micheyl

© Copyright by  
Mounya Elhilali  
2004

## ACKNOWLEDGMENTS

Many dreams can only come true with the help of so many people, and it is simply impossible to say thank you for making dreams come true. This dissertation and entire ph.d experience is one of those dreams. Their very foundation is sustained by some of the most wonderful people. For those of you named, and those inadvertently missed, I owe my heartfelt gratitude for an experience I will cherish forever.

To Shihab, my advisor and mentor, thank you for the most unique and exciting graduate experience I could have ever imagined. To Dr Jonathan Simon, thank you for your always insightful and interesting discussions. To Drs Liu, Reggia, and Micheyl, thank you for accepting to be part of my thesis committee, and enriching this work with your valuable expertise.

To Ray, a special friend and critic. Thank you for so many big and little things I cannot begin to enumerate. I owe you so much for having to listen to my complains and whining, and for helping me always get perspective on things.

To Jonathan Fritz, thank you for expanding my knowledge and fascination with the world of neuroscience, for a very interesting and rewarding collaboration over the years, and for the pleasures of exciting political debates with Shihab. Special thoughts go also to past and present members of the Neural Systems Laboratory. To Sridhar, Nima, Tai-Chih, Elena, Didier, Nikos, Sudha, Jyoti, Kolo, and everybody else. Thank you for helping me in so many ways, and most of all, thank you for your friendship.

And for helping me remember that life is actually so much more than just work, I

am deeply grateful to all my friends here in the US and back home. To my dear Layla for being there for me all the time, and keeping me up to date with what's going on around me. To Chiraz and Ghada for sharing so many incredible adventures with me. Thank you.

This dissertation is certainly dedicated to my parents, for opening my eyes to the world of science, and teaching me to always shoot for the moon. Thank you. To my brother Zouhair and his little family whose love and encouragement have been a great support for me. Thank you.

College Park, November 2004

*Mounya*

# TABLE OF CONTENTS

<b>List of Figures</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 From sound to meaning . . . . .	2
1.1.1 The magic of our auditory system . . . . .	3
1.1.2 Challenges of audition . . . . .	5
1.2 Thesis outline . . . . .	7
<b>2 Auditory system primer</b>	<b>10</b>
2.1 Nature of sound . . . . .	10
2.2 Auditory pathways . . . . .	12
2.2.1 Auditory periphery . . . . .	12
2.2.2 Central pathways . . . . .	16
2.3 Neural receptive fields . . . . .	20
<b>3 Cortical timing paradox</b>	<b>23</b>
3.1 Exploring cortical dynamics . . . . .	25
3.1.1 The stimulus space . . . . .	25
3.1.2 Cortical response properties . . . . .	30

3.1.3	Temporal resolution of cortical information . . . . .	32
3.2	Modelling cortical responses . . . . .	35
3.2.1	Defining a piecewise linear model . . . . .	35
3.2.2	Cortical functions with dual operating-point . . . . .	36
3.2.3	Cortical receptive fields . . . . .	40
3.3	Neural mechanisms . . . . .	47
3.3.1	Synaptic dynamics . . . . .	48
3.3.2	Cortical circuitry . . . . .	50
3.3.3	Emergence of cortical STRFs . . . . .	51
3.4	Functional significance . . . . .	52
<b>4</b>	<b>Speech Intelligibility</b>	<b>54</b>
4.1	Measures of intelligibility . . . . .	55
4.2	Intelligibility of communication channels . . . . .	57
4.2.1	STMI <sup>R</sup> procedure . . . . .	57
4.2.2	Modulation Transfer Functions in noise . . . . .	61
4.3	Intelligibility for noisy speech . . . . .	63
4.3.1	STMI <sup>T</sup> procedure . . . . .	63
4.3.2	Human psychoacoustic testing . . . . .	67
4.4	Nonlinear speech distortions . . . . .	68
4.5	Conclusion . . . . .	72
<b>5</b>	<b>Auditory scene analysis</b>	<b>74</b>
5.1	Perceptual principles . . . . .	77
5.1.1	Gestalt principles . . . . .	77

5.1.2	Acoustic correlates . . . . .	80
5.1.3	Top-down effects . . . . .	86
5.2	Literature review of CASA techniques . . . . .	87
5.3	Adaptive ASA architecture . . . . .	91
5.3.1	Unsupervised learning . . . . .	92
5.3.2	Adaptive competitive learning . . . . .	93
5.3.3	Model architecture . . . . .	100
5.4	Implementation of the model . . . . .	102
5.4.1	Pre-processing stage . . . . .	102
5.4.2	Multi-scale representation . . . . .	106
5.4.3	Cortical filtering . . . . .	108
5.4.4	Adaptive learning . . . . .	111
5.5	Results . . . . .	113
5.5.1	Streaming effects . . . . .	113
5.5.2	Speech segregation . . . . .	136
5.6	Summary and Discussion . . . . .	142
<b>6</b>	<b>Conclusion</b>	<b>148</b>
6.1	Thesis overview . . . . .	148
6.2	Future prospects . . . . .	151
<b>A</b>	<b>Derivation of predictive learning</b>	<b>152</b>
A.1	Optimizing the learning function . . . . .	152
A.2	Difference to state-space equation . . . . .	153



## LIST OF FIGURES

1.1	Stages of auditory processing . . . . .	4
1.2	Contour visual illusion . . . . .	6
2.1	The auditory periphery . . . . .	13
2.2	Model of peripheral auditory processing . . . . .	14
2.3	The central auditory pathway . . . . .	16
2.4	Model of central auditory processing . . . . .	19
2.5	Neuronal receptive fields . . . . .	21
3.1	Ripple domain decomposition . . . . .	27
3.2	Schematic of the stimulus envelope and fine structure . . . . .	29
3.3	Rasters of A1 responses . . . . .	31
3.4	Model of correlation function . . . . .	33
3.5	Population results for correlation functions . . . . .	34
3.6	Schematic of STRF computation . . . . .	38
3.7	Examples of STRF triplets . . . . .	41
3.8	Relating cortical receptive fields . . . . .	43
3.9	Responses to harmonic TORCs . . . . .	45

3.10	Predicting cortical responses . . . . .	47
3.11	Simulation of effects of synaptic depression . . . . .	49
3.12	Simulation of cortical dynamics . . . . .	51
4.1	Schematic of STMI computation . . . . .	58
4.2	STMI <sup>R</sup> and STI for white noise and reverberation . . . . .	60
4.3	Modulation Transfer Functions in noise . . . . .	62
4.4	STMI <sup>T</sup> and psychoacoustics for white noise and reverberation . . . . .	67
4.5	Intelligibility under phase jitter . . . . .	69
4.6	Intelligibility under phase shift . . . . .	72
5.1	Spectrogram of a sound mixture . . . . .	75
5.2	Gestalt principles . . . . .	77
5.3	Auditory grouping . . . . .	81
5.4	Internal world model . . . . .	95
5.5	Schematic of stream segregation model . . . . .	102
5.6	Pre-processing stages of adaptive learning model . . . . .	103
5.7	Multi-scale auditory representation . . . . .	107
5.8	Streaming of alternating tones . . . . .	116
5.9	Streaming of alternating ripples . . . . .	119
5.10	Streaming in a cycle of 6 tones . . . . .	121
5.11	Streaming of alternating vowels . . . . .	123
5.12	Segregation by capturing interfering tones . . . . .	125
5.13	Segregation of crossing-trajectories . . . . .	127
5.14	Segregation of crossing-trajectories (2) . . . . .	128

5.15 Sine-wave speech . . . . .	131
5.16 Sine-wave speech (2) . . . . .	132
5.17 Capturing tone in mixture . . . . .	134
5.18 Capturing tone in mixture (2) . . . . .	135
5.19 Segregating speech from original utterances . . . . .	138
5.20 Segregating speech from original utterances(2) . . . . .	139
5.21 Segregating speech mixtures . . . . .	140
5.22 Segregating speech mixtures (2) . . . . .	141
6.1 Receptive field plasticity in the auditory cortex . . . . .	151

## Chapter 1

# Introduction

*“The purpose of computing is insight, not numbers”*

Richard Hamming

We perceive the world through our senses, but only hearing can give us the delight of enjoying a nice musical melody. Hearing is the process of discovering objects surrounding us via the sounds they emit. It is the sense by which our brain consolidates the acoustic information from the environment into a perceptual and cognitive experience.

Theories of auditory perception try to elucidate how the auditory system transforms sound energy patterns into useful information about acoustic events in the environment. Research in the area of auditory processing entails two main directions: (1) an experimental approach which addresses the biological foundation of auditory perception and

its psychoacoustical and behavioral manifestations, and (2) a computational methodology which focuses on building engineering systems and theoretical models for sound processing.

On the experimental front, intense work on the physiology of hearing has expanded our knowledge of the mechanisms of auditory processing in the brain. Higher-level processing, specifically at the level of cortex, is of particular interest as it is the station where the organization of acoustic signals into perceptual patterns takes place; hence leading to a representation of the acoustic environment in terms of auditory objects. While much is known about the neural mechanisms underlying the function of cortical structures, many questions remain unanswered, and we are still far from having a complete picture of how acoustic information is consolidated, and how auditory objects emerge. The temporal code of cortex is an exceptionally powerful clue to understanding the role of auditory cortex in hearing. This thesis examines the dynamics regulating the cortical function, and explores its underlying neural mechanisms.

On the theoretical front, both theorists and engineers have tackled many problems pertaining to auditory processing, ranging from models of cochlear sound filtering to real-time communication systems. The significance of these models lies both in their practical relevance in our everyday life, as well as their intellectual contribution to our understanding of audition and the general function of the brain. In this thesis, we exploit our knowledge of the neural mechanisms of cortical auditory processing to formulate models addressing the questions of speech intelligibility, auditory streaming and sound separation.

## **1.1 From sound to meaning**

Acoustic signals are transmitted by physical disturbance of a medium, causing vibration of the eardrum, and ultimately resulting in our experiencing audible sound [148]. Despite its

pervasiveness in our daily life, sound is a physical phenomenon that is hard to understand in common-sense terms [54]. It is a familiar form of physical energy that is difficult to visualize, either literally or conceptually, but the consequences of its presence are readily discernible. A basic understanding of the physics of sound is an important initial step in addressing the general question of auditory perception. Yet, the road from sound to meaning is one that involves the entire apparatus of the auditory neural circuitry, extending from the ear to the brain. In this sense, the simple presence of physical vibrations in our ear is not what makes us *hear*.

### 1.1.1 The magic of our auditory system

While seemingly effortless, the auditory system performs an incredibly complex task of sound perception. We are equipped with an amazing computational tool that is both competent and quite reliable in perceiving sounds. When listening to the environment around us, sounds from all sources are combined together into one complex auditory field that we have to “navigate” our way through in order to identify the individual sound elements and sources. We do not have separate “pipelines” for each sound object in the environment, as originally thought by ancient Greeks [148]. Sounds in our environment are all lumped together in one acoustic input that reaches our ears. The nervous system takes on the extraordinary job of telling us which instrument is playing in the orchestra, whether a chorus is accompanying the music, and which melody is being played. It also performs its task with an impressive degree of reliability, even in the presence of the most severe distortions. Our ability to follow a conversation carried in a very noisy environment is a testament to how robustly the biology carries its job of *audition*.

The perceptual capabilities of the auditory system rely on various cognitive princi-

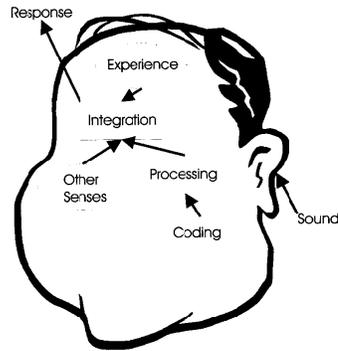


Figure 1.1: Schematic of stages of auditory processing (from [148]).

ples allowing us to attend to certain aspects of the acoustic information in the auditory scene. Such capabilities make us very adept at listening to one voice in the midst of other conversations and noise. While disregarding information from all other surrounding sounds and noise interferences, we are able to orient our attentional focus towards the voice of the speaker.

In the 19<sup>th</sup> century, the Russian writer Valdimir Odoevsky recounted the story of a malicious sorcerer who gave a man the power to hear and see everything. This “gift” caused this unfortunate man to experience nature as a completely fragmented world. Nothing formed into a compact unit in his mind, and sounds became a torrent of erratic meaningless mechanical vibrations, incoherent and with no meaning. This story told by Odoevsky, a musician himself as well as a writer, alludes to how critical it is for the auditory system to be able to attend to particular sound elements in the environment, and to recognize unitary sound streams based on their common physical characteristics and acoustical structures. Sound segregation is in effect one of the most important tasks carried by the auditory system.

The segregation and perception of auditory objects in the environment is achieved through a complex process of sound analysis and cognitive integration (Figure 1.1). The

acoustic environment is parsed into a neural code representing its physical characteristics. The particular structure of this neural code carries the information about the sound signal, which then gets interpreted by the brain, and integrated with other sensory information and prior experiences leading to our perception of the environment around us, and hence to behavioral responses.

### 1.1.2 Challenges of audition

If sensory information is to be useful for the control of behavior, the nervous system must be capable of making reliable perceptual judgments in a very rapid fashion. It must have the capability of processing complex and constantly changing sensory information in “one trial”. This task is not always made easy by nature as sensory information can and does sometimes present conflicting or limited evidence about the environment. Nevertheless, the auditory function adheres to a set of perceptual principles that allow it to organize acoustic information in perceptually meaningful events, despite the lack of sensory evidence. These universal rules are strongly invoked in cases of perceptual illusions, i.e. situations where a normally-functioning brain perceives “things” that may not be physically present. By definition, illusions are perceptual phenomena that emerge as a consequence of psychological principles which do not necessarily reflect an accurate representation of the sensory information. In the case of audition, auditory illusions make us hear things different from the actual nature of sound in the environment. To illustrate such effect, we cite an example from vision, as similar principles underly the function of auditory and visual perception. The phenomenon of *subjective contours* is a well known visual illusion, and is illustrated in Figure 1.2. The figure demonstrates how our visual system perceives contours that are not physically present, in a similar way that the auditory system can



Figure 1.2: Contour visual illusion (from [137]).

perceive sounds that are not there.

Just like vision, auditory computation is a dynamic process. It is capable of constructing precise representations of the world by complementing signals in the environment with information about the global context of an auditory scene. Internal models of the world and prior knowledge and experience reflect our expectations from sensory inputs and adapt to the changing flow of information coming from the environment, while adhering to a universal set of perceptual rules. These intricate interactions complicate our job of studying complex systems such as auditory processing, and compel us to presuppose certain simplifications about the elements and structures of the system. It is not clear, however, which level of abstraction or modelling assumptions are more appropriate for the study of auditory perception, leading to a real controversy concerning the best approach for modelling biological systems.

Looking back at the history of the computational theory of perception, an early and important contribution was made by Gibson in the 1960s. His work is one of the early attempts to understanding complex information-processing systems. According to Gibson, “*the function of the brain, when looped with its perceptual organs, is not to decode signals, not to interpret messages, not to accept images, not to organize the sensory input or to process the data, in modern terminology. It is to seek and extract information about the environment from the flowing array of ambient energy*” ([102], page 29). Gibson and others were later criticized for their oversimplification of complex information processing

in the brain and their *simplistic* understanding of perception. Marr wrote that Gibson “*did not understand properly what information processing was, which led him to seriously underestimate the complexity of the information-processing problems ...*” [102]. Later on, Marr himself was criticized by Churchland and colleagues [35] who argued that his pure vision view of the world is itself a dangerously simplified caricature of the problem of visual perception. Slaney raised similar concerns about the simplified and purely bottom-up views of perception in the auditory field [127].

As theoretical debate still continues about the proper direction for studying complex sensory processing in the brain, we adopt a middle-ground approach in our study of auditory processing. While mainly concerned with the neural encoding of sound in the auditory system in a primarily “bottom-up” direction, we also consider aspects of “higher-level” influences, particularly in addressing aspects of auditory processing related to scene analysis.

## 1.2 Thesis outline

As we describe the intricacy of the hearing problem and its multiple facets, our sense of the complexity of tackling its different aspects is only reinforced. In this thesis, we focus on the computational task of hearing by raising questions such as which sensory information does the brain extract, and how is it extracted and processed. This approach views the auditory system as a computational tool that is separate from the anatomy in which it is implemented. Nonetheless, as we focus on the higher-level functions of auditory processing in the brain, it is only natural to explore its neural basis, and particularly at the cortical level. By setting a biological foundation for the computational problem, we can explore different implementations and modelling schemes targeting specific questions

of sound processing.

This dissertation is organized in six chapters. Following this introduction is an auditory system primer. In that chapter, we briefly review the present knowledge of the biological auditory system. The survey summarizes what we know about the organizational and functional structures of the auditory system, focusing on the principles that are essential in understanding how the computational models proposed in this thesis relate to the biology. We also touch upon some unresolved questions concerning the role ascribed to the auditory cortex, and we dedicate the next chapter to expand on these issues.

Chapter 3 explores evidence about known paradoxical properties of cortical neural processing, namely the dynamics that underly temporal coding of sound signals in the auditory cortex. We analyze neural data collected in the primary auditory cortex, and argue for a dual role of cortical function in organizing the features of sound, in terms of both slow spectrotemporal information patterns (syllabic segments in speech, timbre and rhythm in music); as well as more fast transient and precise responses capturing the sound texture. This study has important implications in the way we understand how auditory percepts are formed in the brain.

Chapter 4 presents a computational approach for evaluating loss of intelligibility, inspired from mammalian physiology and human perception. The model is based on a multi-resolution filter-bank implementation of cortical response patterns. A powerful and robust *computational* intelligibility measure based on this model is presented, where estimates of the integrity of spectrotemporal modulations in a test signal or channel are related to *perceptual* measures of speech intelligibility, as perceived by humans.

In Chapter 5, we explore another aspect of auditory perception, that of auditory scene analysis and stream segregation. We develop and test a cortical model for sound

organization based on adaptive learning and Kalman estimation. The model is founded on perceptual principles of auditory grouping and stream formation. Such principles are translated in a computational model which combines aspects of bottom-up sound processing with an internal representation of the world, which adapts its intrinsic representation based on the residual error between its own predictions and the actual sensory input. The model proves to be quite powerful in organizing sounds in perceptual streams that correspond to the actual perceived events in real-life situations. We present various simulations addressing different aspects of auditory streaming, as well as sound separation from speech mixtures.

Finally, we conclude in Chapter 6 with a summary of the main findings of this work, and consider further prospects of this research field.

## Chapter 2

# Auditory system primer

Hearing is one of the means by which organisms determine objects in their environment. If these objects vibrate, they have the potential to produce sound, and that sound can be an identifying characteristic of the object [66]. This chapter is intended to review our current knowledge of the auditory system, and how it processes sound. We limit this survey to the basic elements of biological auditory processing that are necessary for the purpose of this thesis. Some existing computational models are also briefly reviewed as they are relevant for our subsequent analysis of models for auditory processing.

## 2.1 Nature of sound

It is appropriate to start our study of sound processing by a description of the physical properties of sound itself. Any object that vibrates can produce an audible sound, as

a form of energy whose behavior is governed by the laws of physics. Sound is in fact a physical disturbance that propagates through any elastic medium. It is hence the air vibration -rather than the sound source itself- that starts the hearing process [66].

Sound is captured by an instantaneous amplitude or pressure waveform that varies in time, making the latter the first dimension of hearing. As de Cheveigné puts it so well [44], “*Time must flow for sound to exist*”. There is strong physiological evidence that the time-variability of the acoustic waveform is encoded either explicitly or implicitly in the temporal patterns of neural responses. These temporal cues are parsed by the auditory system to identify various sound properties such as source, identity, location and meaning. Patterns representing different sound properties extend over a large range of time scales, extending from microseconds (for sound localization) to hundreds of milliseconds (syllabic segments) and even many seconds (phrase duration). The auditory system is hence responsible for resolving information at different temporal resolutions and integration windows.

Sound can be represented equally well by its frequency content. The Fourier theory transforms the functional dependence of a signal from time to frequency [109]. It reveals the frequency attributes of sound, and hence offers an equivalent representation of any sound in terms of its individual frequencies (or sinusoidal vibrations). While both time and frequency are valid dimensions for representing sound, the dynamic changes of sound requires in effect a short-term analysis of its frequency content as the signal varies over time. Biology seems to be “aware” of this fact, and is able to put together these two representations (time and frequency) by mapping the spectral axis into a spatial axis, making frequency the second dimension of hearing. This spatial axis is represented by

the tonotopic<sup>1</sup> organization of the cochlea as well as the tonotopic maps found in various auditory nuclei, as we shall see next in our review of the auditory pathway.

## 2.2 Auditory pathways

The hardware of the auditory system consists of the ear and parts of the central nervous system (CNS). Technically, the term *ear* refers to the entire peripheral auditory apparatus including the outer, middle, and inner ear. Sound information is then projected along a multitude of channels making up the main auditory pathway. In the following section, we briefly review the structure of the different auditory nuclei and review basic knowledge about their role in sound perception.

### 2.2.1 Auditory periphery

Incoming sound waves entering the ear make the eardrum vibrate. The sound energy is then converted into mechanical energy, which produces a complex spatio-temporal pattern of vibrations along the basilar membrane of the cochlea (Figure 2.1). The maximal displacement at each cochlear point corresponds to a distinct tone frequency in the stimulus, creating a tonotopically ordered response axis along the length of the cochlea [54, 148]. The basilar membrane can then be thought of as a mechanical short-term Fourier analyzer of sound frequency [87].

Following the cochlear stage, the stimulus frequency and intensity are then encoded at the level of the cochlear nerve fibers through innervation of the inner hair cells. While the activity of the nerve fibers is thought to be relatively homogeneous in terms of vibration

---

<sup>1</sup>The term “tonotopy” is used in the auditory literature to refer to the organization of frequency along a logarithmic spatial axis, much like a xylophone. It must be noted, however, that the tonotopic organization in the mammalian auditory system is not purely logarithmic, but tends to become linear below 500 Hz [81].

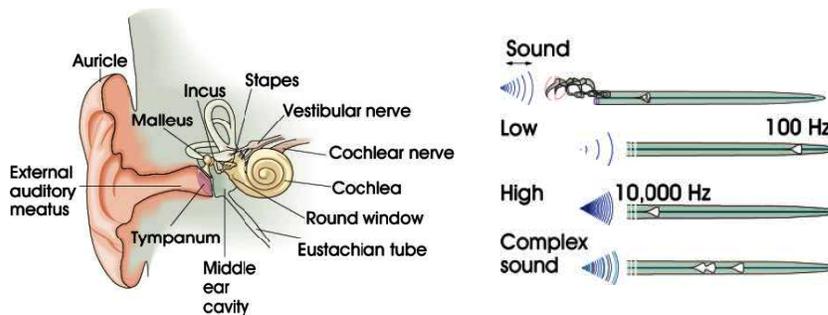


Figure 2.1: Structure and functionality of the human ear [87]. **(left)** Vibrations to the eardrum are conveyed across a fluid-filled middle ear by three tiny, linked bones. These in turn stimulate the cochlea by producing oscillatory pressure causing a travelling wave along the basilar membrane. **(right)** Each frequency excites maximal motion at a particular position along the basilar membrane, allowing the cochlea to perform a spectral analysis of sound.

patterns [109], the next relay station -the cochlear nucleus, appears to be more intricate. Various projections within the cochlear nucleus constitute parallel pathways for analyzing different sound attributes. Present evidence suggests a role of the cochlear nucleus in enhancing and sharpening the features of the neural patterns, prior to relaying them to more central areas via the superior olivary complex (SOC) and the inferior colliculus (IC) [116].

### Computational model

To mimic the functionality of peripheral auditory processing, we use a computational model that is grounded on extensive neurophysiological data from mammalian peripheral stages of auditory processing [100, 147]. The choice of this model is motivated by its biological foundation, its perceptual relevance as well as noise robustness as shown by thorough analytical and experimental investigations established by Wang *et al.* [144]. It is used in this thesis as one of the building blocks for our subsequent analysis of compu-

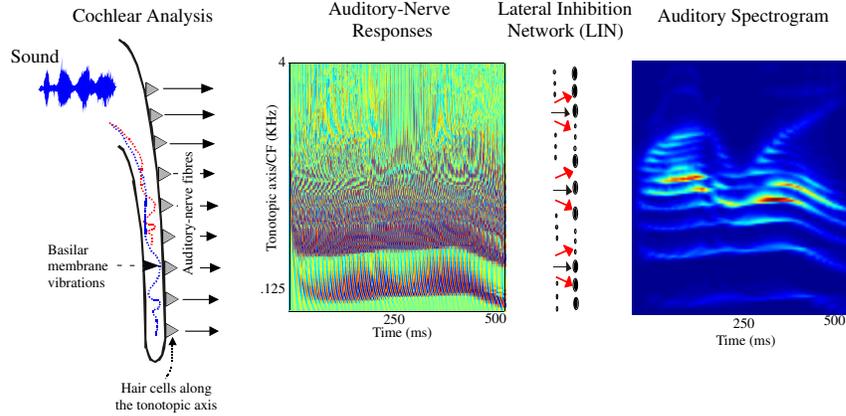


Figure 2.2: Schematic of the early stages of auditory processing. Sound is analyzed by a model of the cochlea (depicted on the left) consisting of a bank of 128 constant-Q bandpass filters with center frequencies equally spaced on a logarithmic frequency axis (tonotopic axis) spanning 5.2 octaves (e.g., 0.1-4kHz). Each filter output is then half-wave rectified and lowpass filtered by an inner hair cell model to produce the auditory-nerve response patterns (middle panel). A spatial first-difference operation is then applied mimicking the function of a lateral inhibitory network (LIN) which sharpens the spectral representation of the signal and extracts its harmonics and formants [131]. The short-term integration is typically performed over 8 ms intervals. A final smoothing of the responses on each channel results in the auditory spectrogram depicted on the right.

tational strategies of auditory perception. In this section, we describe briefly the steps involved in computing an *auditory spectrogram* based on the original work presented in [144, 147]. While not strictly biophysical, the model abstracts from physiological data relevant for basic sound analysis. It consists of various stages based on a wavelet-analysis of the acoustic waveform ( $s(t)$  in Equation 2.1), modelled as a three-step process:

- First, the frequency analysis in the cochlear stage is modelled by a bank of constant-Q highly asymmetric bandpass filters ( $Q=4$ ) equally spaced on a logarithmic frequency axis ( $h(t, x)$  in Equation 2.1). The model employs 24 filters/octave over a 5.3 octave range. The left panel of Figure 2.2 illustrates an incoming sound waveform processed through a bank of frequency selective filters.

- Next, the basilar membrane outputs are converted into inner hair cell intra-cellular potentials ( $y_2(t, x)$  in Equation 2.1). This process is modelled as a three-step operation: a high-pass filter (the fluid-cilia coupling), followed by an instantaneous nonlinear compression (gated ionic channels) captured by a nonlinear function  $g(\cdot)$ , and then a low-pass filter  $w(\cdot)$  (hair cell membrane leakage). Detailed description of the mechanisms involved in each one of these steps can be found in [100].
- Finally, a lateral inhibitory network detects discontinuities in the responses across the tonotopic axis of the auditory nerve array, inducing a sharpening of the filter-bank frequency selectivity as seen in the cochlear nucleus [131]. It is modelled as a first difference operation across the channel array, followed by a half-wave rectifier, and then a short-term integrator. The temporal integration window is captured by the function  $\mu(t; \tau) = e^{-t/\tau}u(t)$  with time constant  $\tau$ . This stage effectively sharpens the bandwidths of the cochlear filters from about  $Q=4$  to 12, as explained in detail in [144].

$$\begin{aligned}
y_1(t, x) &= s(t) *_t h(t; x) \\
y_2(t, x) &= g(\delta_t y_1(t, x)) *_t w(t) \\
y_3(t, x) &= \max(\delta_x y_2(t, x), 0) *_t \mu(t; \tau)
\end{aligned} \tag{2.1}$$

Effectively, the above sequence of operations computes an auditory spectrogram of the acoustic signal (Figure 2.2) using a bank of constant-Q filters, with a bandwidth tuning  $Q$  of about 12 (or just under 10% of the center frequency of each filter). Dynamically, the spectrogram also explicitly encodes all temporal “envelope modulations” in the signal due to interactions between the spectral components that fall within the bandwidth of each filter. The frequencies of these modulations are naturally limited by the maximum bandwidth of the cochlear filters.

## 2.2.2 Central pathways

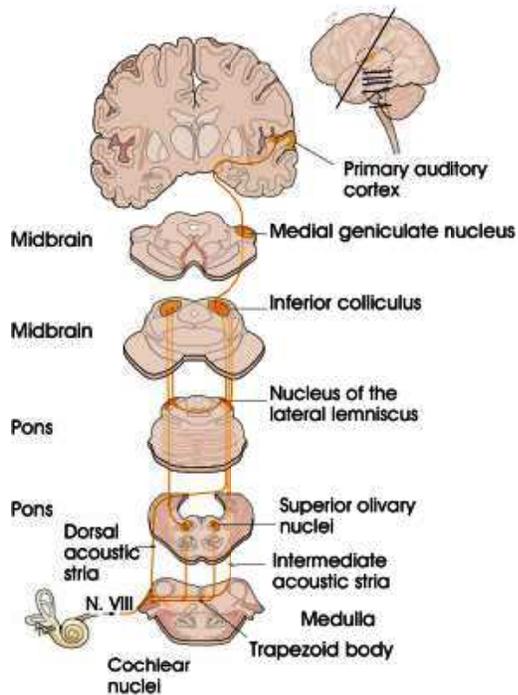


Figure 2.3: The central auditory pathway [87]. The central pathways extend from the cochlear nucleus to the auditory cortex through a complex midbrain circuitry. Each nucleus mediates certain functions, such as sound localization and binaural sensory integration. It is important to note that the projections along the auditory pathway are far more complex than those shown in the figure. For instance, there are extensive feedback loops from the auditory cortex into the thalamus and inferior colliculus, which introduce highly non-linear dynamics into the system [11].

At this stage along the auditory pathway, the acoustic waveform is converted into a pattern of neural activity that faithfully maps the temporal and spectral attributes of the incoming sound. The following stages begin the process of temporal integration and the formation of a coherent auditory image. Nuclei of the midbrain, the Inferior Colliculus (IC) in particular is believed to to be the first stage in the system where all the acoustic information converges together, coming through projections from the superior olivary complex and the lateral lemniscus [87] (Figure 2.3). The IC appears to act both as

an integrative station as well as a switchboard to higher auditory and multi-modal sensory areas. It also plays a key role in binaural hearing and thus in sound localization [55].

If the IC is believed to be the last station in the auditory pathway whose function is directed towards the formation of auditory images, the subsequent auditory nuclei are thought to play a role in the analysis of this auditory image and hence in the perception of sound [55]. The nuclei beyond IC, including the auditory thalamus (Medial Geniculate Body, MGB) and auditory cortex are hence involved in the process of auditory pattern recognition, where the features of the auditory image are sharpened and grouped together into streams, mediated by auditory memory and contextual information. Present evidence seems to favor a role of cortical circuitry in auditory pattern recognition. As most of the interesting auditory features are already extracted by the level of the IC, it is suggested that the auditory cortex is playing a role in organizing these features in terms of auditory objects [112].

It is, however, very important to stress that the structure and function of the central auditory nervous system is far less understood than the periphery. The anatomical complexity of the pathways, the neural morphology of cells and circuitry, as well as the unknown nature of the neural code have made it very difficult to study the central pathways of the auditory system. Our current knowledge of the auditory functions mediated by the different nuclei is very limited, and greatly speculative. Nonetheless, evidence from imaging as well as physiological and anatomical studies are laying the grounds for a better understanding of the function of the central auditory system and the neural strategies of sound perception.

## Computational model

Despite claims that feature extraction is a secondary role of cortical circuitry in sound processing [113], the case for the real contribution of cortex to auditory perception is far from being resolved. The *simplistic* view of cortical neurons as “feature detectors” is by itself not erroneous. We will show in this thesis how we can build on a simple view of A1 as a feature detector to perform elaborate auditory functions, by taking into account additional physiological facts to enhance our understanding of the cortical function. We shall expand further on this topic in chapter 3.

Inspired from the computational role of cortical circuitry in auditory pattern recognition, we adopt a model presented by Chi *et al.* [32] to mimic the functionality of central auditory processing. The basis for this model is derived from physiological data in animals [49, 90, 91], and psycho-acoustical data in humans [31].

The model consists of a multi-scale filter-bank represented by impulse responses in the form of spectrotemporal Gabor functions [31]. Each filter is tuned to a range of temporal (denoted  $\omega$ , or rate) and spectral (denoted  $\Omega$ , or scale) modulations, with a spatial impulse response  $h_{\mathcal{RF}}(x; \Omega_c, \phi_c)$  and temporal impulse response  $g_{\mathcal{RF}}(t; \omega_c, \theta_c)$ . The overall impulse response of each filter is then a “separable” spectrotemporal modulation function  $\mathcal{RF}$ , given as the product of a temporal and spectral marginal functions (middle panel of Figure 2.4):

$$\begin{aligned}
 g_{\mathcal{RF}}(t; \omega_c, \theta_c) &= g(t; \omega_c) \cos \theta_c + \hat{g}(t; \theta_c) \sin \theta_c \\
 h_{\mathcal{RF}}(x; \Omega_c, \phi_c) &= h(x; \phi_c) \cos \phi_c + \hat{h}(x; \phi_c) \sin \phi_c \\
 \mathcal{RF}(t, x; \omega_c, \theta_c, \Omega_c, \phi_c) &= g_{\mathcal{RF}}(t; \omega_c, \theta_c) \cdot h_{\mathcal{RF}}(x; \Omega_c, \phi_c)
 \end{aligned}
 \tag{2.2}$$

This multi-scale multi-rate representation simulates the selectivity of cortical neurons to

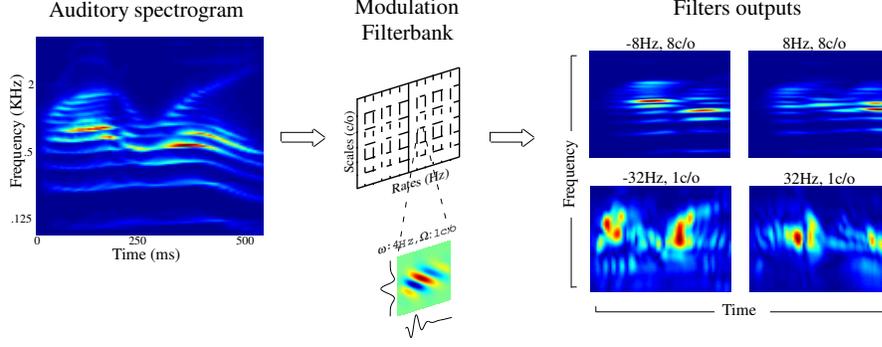


Figure 2.4: The cortical multi-scale representation of sound. The auditory spectrogram of a speech sentence /right away/ (from Figure 2.2), spoken by a male is analyzed by a bank of spectrotemporal modulation selective filters. The spectrotemporal response field (STRF) of one such filter (tuned to  $\omega = 4$  Hz and  $\Omega = 1$  cycles/octaves) is shown in middle panel. The output from each filter is computed by convolving the STRF with the input spectrogram, to produce a new spectrogram as shown in the right panels. The panels show the magnitude response of 4 such filters.

spectral local shapes, rate movements of spectra, as well as direction of movement (upward or downward) (right panels of Figure 2.4). The spectrotemporal response of each filter to an input spectrogram  $y(t, x)$  is given by:

$$\begin{aligned}
 r(t, x; \omega_c, \theta_c, \Omega_c, \phi_c) &= y(t, x) *_{xt} \mathcal{R}\mathcal{F}(t, x; \omega_c, \theta_c, \Omega_c, \phi_c) \\
 &= y(t, x) *_{xt} [g_{\mathcal{R}\mathcal{F}}(t; \omega_c, \theta_c) \cdot h_{\mathcal{R}\mathcal{F}}(x; \Omega_c, \phi_c)] \\
 &= y(t, x) *_{xt} [g \cdot h \cos \theta_c \cos \phi_c + g \cdot \hat{h} \cos \theta_c \sin \phi_c \\
 &\quad + \hat{g} \cdot h \sin \theta_c \cos \phi_c + \hat{g} \cdot \hat{h} \sin \theta_c \sin \phi_c]
 \end{aligned} \tag{2.3}$$

The output can be reduced to a 4 dimensional complex-valued mapping obtained from a complex valued wavelet transform varying along time, frequency, spectral scale, temporal rate. A functional description of the parameters of the cortical model is presented in [145].

Figure 2.4 illustrates the analysis stages through the multi-scale filter-bank. The input spectrogram is decomposed through the various filters into a four-dimensional complex-

valued response (time, frequency, rate, and scale). Different views of this response can be obtained by summing the outputs of all filters along one or more dimensions. The magnitude of the output of 4 modulation selective filters are shown in the right panels of Figure 2.4. The fast filters (+32Hz,-32Hz) reflect the fast temporal envelope in the original speech utterance viewed through a 32 Hz filter, while the slow envelope dynamics are captured by the overall patterns of the 8Hz filter. We can also note the different patterns in the upward vs. downward filter responses, demonstrating the orientation selectivity of the cortical filter-bank model. The spectral patterns can also be observed at the output of the 1 vs. 8 cycles/octave.

### 2.3 Neural receptive fields

As we try to understand how the world is represented in the brain, a natural step is to characterize the selectivity of sensory neurons to external stimuli, since the main workload of information processing in the brain is carried out by neurons [87]. Clearly, different neurons respond variably to different stimulus patterns, and a functional description of each cell's behavior is its *receptive field*. A receptive field is a description of the optimal input that elicits the strongest response in a neuron. In the case of auditory neurons, the receptive field is generally described as a two-dimensional (spectral and temporal) functional, called STRF (Spectro-Temporal Receptive Field), which acts as a time-dependent spectral transfer function, or a frequency-dependent dynamical filter [47, 88, 138].

In general, biological systems fall in the category of dynamical systems whose input-output transformation can be described by a possibly nonlinear dynamical functional  $\mathcal{F}[\cdot]$  mapping the values of the input  $s$  at different time instants to a value of the output or response  $r$  at the current time  $t$  [89]. Under assumptions of bounded input amplitude,

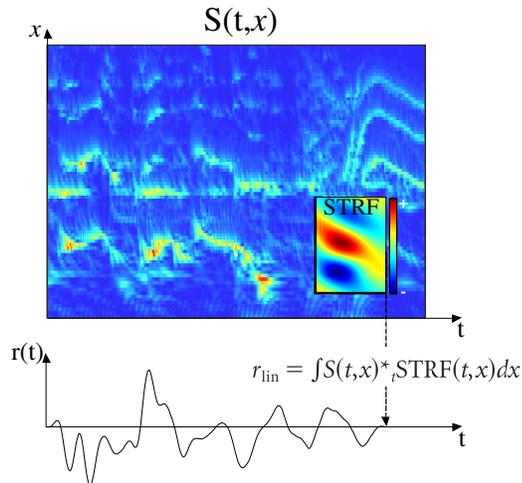


Figure 2.5: A receptive field captures the optimal input that elicits the strongest response in a neuron. The convolution of the receptive field with the spectrotemporal representation  $S(t, x)$  of a stimulus captures the linear component of the response of the neuron  $r(t)$ .

the functional  $\mathcal{F}$  can be expanded into Volterra series components, as described by the equation:

$$r = \mathcal{F}[s] = \mathcal{V}_0[s] + \mathcal{V}_1[s] + \dots \mathcal{V}_n[s] \quad (2.4)$$

where,

$$\mathcal{V}_i[s(t)] = \int \dots \int v_i(\tau_1, \dots, \tau_i) \cdot s(t - \tau_1) \dots s(t - \tau_i) d\tau_1 \dots d\tau_i, \quad (2.5)$$

and the order  $n$  could be conceivably infinite. A Volterra expansion is a generalized Taylor series for nonlinear dynamical systems [30]. Using this general framework of system representation, the receptive field of neurons is nothing but a second-order Volterra kernel ( $v_2$ ).  $v_2$  describes a linear system transforming the time-dependent autocorrelation of the stimulus (or equivalently, a time-frequency representation of the stimulus), to the response  $r(t)$  [88]. This kernel is named spectrotemporal receptive field (STRF) (Figure 2.5).

STRFs have indisputably revealed a lot about the behavior of neurons in the auditory system, especially at the cortical level. They offer a straightforward quantitative

linear description of the neurons' selectivity to specific stimulus patterns, and hence led us to a better understanding of cortical processing. They have, however, been criticized from many shortfalls:

- *Linear model:* When probed with more behaviorally relevant natural sound ensembles, the linear STRF model proves to be an incomplete description of response properties of nonlinear auditory neurons [138], and fails to successfully predict responses to many natural stimuli. A success prediction rate of 10% [101] to 40% [128] was typically reported for classes of natural sounds.
- *Lack of generalization:* While the STRF model seems to give satisfactory results for a large set of stimulus ensembles (ripples, modulated noise, random tone pips, classes of natural sounds, etc), it appears that comparisons of receptive fields obtained from different stimulus bases leads to striking difference between the derived kernels [138]. STRFs have also been reported to lack robustness relative to stimulus perturbations, such as use of background noise with natural stimuli [9].
- *Describing cortical dynamics:* STRF models fail to capture various aspects of cortical processing. If anything, they reveal a paradox in cortical dynamics, suggesting a slow time constant for cortical processing, despite the known ability of A1 neurons for temporal precision. For instance, STRF models cannot explain the sensitivity of A1 neurons to certain acoustic patterns, such as fast FM sweeps [113].

We shall explore some of these limitations in the following chapter, in our quest for a better understanding of the auditory function and perception of sound.

## Chapter 3

# Cortical timing paradox

The current knowledge of central auditory processing raises many questions related to the neural basis of auditory perception and the physiological foundation of auditory scene analysis. The involvement of cortical circuitry in representing sounds as auditory objects remains to be investigated. Recent evidence suggests a role of cerebral cortex in scene analysis and auditory object formation (see review by Nelken [112], and references therein), but the dynamics of cortical processing have not been carefully addressed in the literature. To quote Nelken [113], *“The issue of time constants is crucial for understanding processing in A1... Studying the interplay between the different time constants will lead us to a better understanding of the operations performed by A1 and therefore to a more precise formulation of its role in the auditory pathway”*. In this section of the thesis, we try to address some of these very critical yet unanswered questions. At what time

constants do cortical neurons operate? Are these time constants (or integration windows) commensurate with the temporal dynamics of stream formation and auditory grouping?

The physiological evidence accumulated over many decades of research shows that temporal responses in auditory neurons are surprisingly sluggish [39, 90, 107]. The upper-limit of sustained locking to repetitive stimuli in A1 does not generally exceed 20 – 30 Hz [130]. Yet, the primary auditory cortex has been shown to be capable of remarkable temporal precision. Various studies have shown that A1 neurons respond very accurately to sound onsets and rapid transients with precision of the order of few milliseconds [5, 50, 51, 79]. Similar findings have also been reported in other sensory systems such as the visual [8], and somatosensory cortex [122]. This apparent paradox in time scales has a perceptual manifestation in the so-called *resolution-integration paradox* [48, 141]. Originally put forward by deBoer [46], this paradox addresses the contradiction in the system’s role in integrating information over long periods of time, and yet being able to maintain a rapid response and fine temporal resolution. Simply put, how is it possible for a system to be both slow and fast at the same time?

Our main interest for raising such questions is to explore the computational strategies that govern the cortical function. To do so, we re-examine physiological data collected from cortical neurons. We start by summarizing our basic findings concerning the accuracy and extent of precise spiking in the primary auditory cortex. Next, we compare receptive fields derived from the envelope and fine structure of the stimulus, and explore their relationship and their ability to account for the details of cortical responses. Finally, we examine whether synaptic depression and specific excitatory/ inhibitory mechanisms can account for these findings, and the possible functional relevance of the fine structure in auditory perception. The work presented here has been published in [61, 62].

## 3.1 Exploring cortical dynamics

In addressing the resolution-integration paradox, researchers have generally studied these two phenomena separately using stimuli that tend to highlight one or the other. For instance, cortical responses were entrained using amplitude and frequency-modulated tones and noise, drifting gratings, and click trains [56, 99, 130], whereas transient responses were evoked using tone onsets and dynamic dots [8, 79]. In this work, we investigate the coexistence of these two response properties in single-units of the primary auditory cortex (A1), and explore their limits and characteristics with stimuli that combine both repetitive and transient features. Understanding the interplay between these time constants leads to a better understanding of the role of cortical circuitry in perceiving and grouping sounds.

### 3.1.1 The stimulus space

#### Time-frequency space

The nature of sound is set by its *conjoint* spectral and temporal attributes. Very often, physiological investigations have assumed these two dimensions to be processed independently, and hence spectral response fields [129] and rate tuning curves [91] have been used as neural descriptors. It is becoming increasingly clear that the combination of *both* spectral and temporal sound properties, and not simply their individual attributes, is important for auditory perception. The spectrotemporal domain is hence a natural choice as a stimulus space. Not surprisingly so, as spectrotemporal representations are inherent to the auditory system both anatomically and functionally.

We can best describe the time-frequency space in terms of *Fourier series* [118], as this latter is a natural analytic description of any dynamic spectrum. It works by decomposing a given sound spectrum into its constituent elementary 2D Fourier components, as

illustrated in Figure 3.1. Each Fourier element is a sine wave that is a function of both time  $t$  and spatial location  $x$ , and is described as:  $A\cos\{2\pi(\omega t + \Omega x) + \psi\}$ . The spatial axis  $x$  corresponds to logarithmic frequency, i.e.  $x = \log_2(f/f_0)$ , where  $f_0$  is the lowest frequency component of the signal. The elements  $\omega$  (in Hertz, or cycles/second) and  $\Omega$  (in cycles/octave) are referred to as temporal and spectral modulation frequencies (since they reflect the modulation energy in the acoustic spectrum at the specific frequencies  $\omega$  and  $\Omega$ ). Each individual Fourier component, termed *ripple*, is then associated with a unique frequency modulation pair  $(\omega, \Omega)$  and defined by its peak amplitude  $A$  and phase  $\psi$  (Figure 3.1 (A)). Just like pure tones (or single sinusoids) can represent any finite-duration acoustic waveform, the 2D ripples can similarly decompose any acoustic spectrum of finite duration and finite bandwidth into a unique set of ripple components. Being a complete basis set, the decomposition of any signal  $S(t, x)$  is captured by its ripple content:

$$S(t, x) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} a_{k,l} e^{j[2\pi(\omega_k t + \Omega_l x) + \psi_{k,l}]} \quad (3.1)$$

### Stimulus parameters

Conventionally, cortical circuitry has been thought to encode sound envelopes. Previous studies [47, 65, 108, 138], including work from the Neural Systems Laboratory [49, 90, 88], tried to characterize cortical responses in terms of their sensitivity to edges and patterns of the stimulus profile or envelope. Receptive fields of neurons (STRFs) have been derived with a variety of stimuli: simple tone pulses [47], animal vocalizations [138], natural sounds [101], white noise [80] and dynamic ripple [49, 108]. Irrespective of the stimulus choice or parameter space (filter-bank output [138], spectrotemporal envelope [65, 88], Wigner distribution [58]), these studies have all used some form of representation of the sound

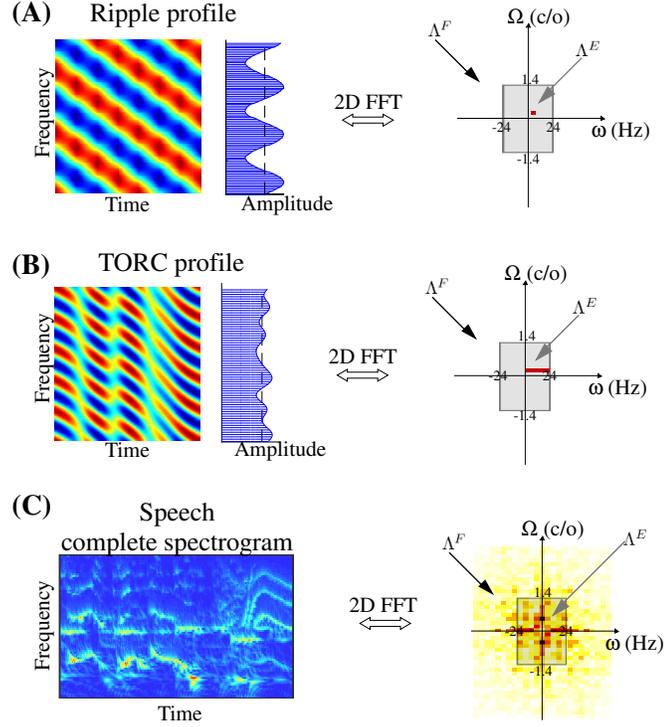


Figure 3.1: Schematic of stimulus spectrotemporal space. **(A)** A ripple profile with modulations  $\{4\text{Hz}, 0.4 \text{ cycles/octave}\}$  appears as a single element in a 2D Fourier transform (red square in the right panel). Ripples are basis functions for the spectrotemporal space. **(B)** TORC stimuli have a more rich envelope (leftmost panel), but is only confined within the  $\Lambda^E$  space. We only show the Fourier decomposition of the TORC profile, but clearly the noise carrier (middle panel) contributes to the TORC energy outside the  $\Lambda^E$  region. **(C)** A “complete” spectrogram of a general sound fills in the entire spectrotemporal Fourier space (rightmost panel).

spectrum (or dynamic envelope) in the time-frequency space.

In contrast, the current study tries to use the entire time-frequency space described above as working ground for defining the stimulus set. It expands the stimulus representation beyond the envelope profile of the stimulus, to also comprise the fast dynamics of the sound, including the interplay between the envelope and the carrier of the stimulus. We therefore formalize the description of the stimulus in terms of two main regions in the time-space domain:

- *The stimulus profile:* The stimulus profile is a steadily drifting spectral envelope, moving at various slow rates [61, 88]. This stimulus, called TORC (Temporally-Orthogonal Ripple Combination), consists of a linear sum of individual ripple elements with  $\omega_k \in \{4, 8, \dots, 24\}$  Hz, and  $\Omega_l \in \{\pm 0.2, \pm 0.4, \dots, \pm 1.4\}$  cyc/oct. These two parameters define the region  $\Lambda^E$  in spectrotemporal space (Figure 3.1). This region spans a range of spectral and temporal modulations previously shown to elicit phase-locked responses in A1 [90, 91]. Sounds extended over 1 to 3 sec in time with a periodicity  $T = 250\text{ms}$ , and spectral bandwidth  $X = 5$  octaves.
- *The stimulus fine structure:* Beyond region  $\Lambda^E$ , the stimulus is defined by its noise carrier, which once transduced through the cochlear hair cells, creates frequency beatings that mostly define the content in region  $\Lambda^F$  (Figure 3.1). The region  $\Lambda^F$  is harder to formalize by design, since it depends on a realization of white noise with random-phase tone components, as well as the mechanics of cochlear filtering. In this study, we use the model described in chapter 2 to mimic as closely as possible the effect of biological peripheral processing. The nature of the noise carrier used in this study came in two variants:
  1. A white noise carrier consisting of 501 random-phase tones, equally-spaced along the tonotopic frequency axis, and spanning a range of 5 octaves, or,
  2. a broadband carrier with harmonically-spaced tones, also spanning a 5 octaves range. The harmonic fundamental frequencies used in the experiments spanned the range (25-200) (Hz).

Throughout this study, all stimuli shared a common (logarithmic or harmonic) carrier consisting of the same instance of frozen broadband noise. As mentioned earlier, a

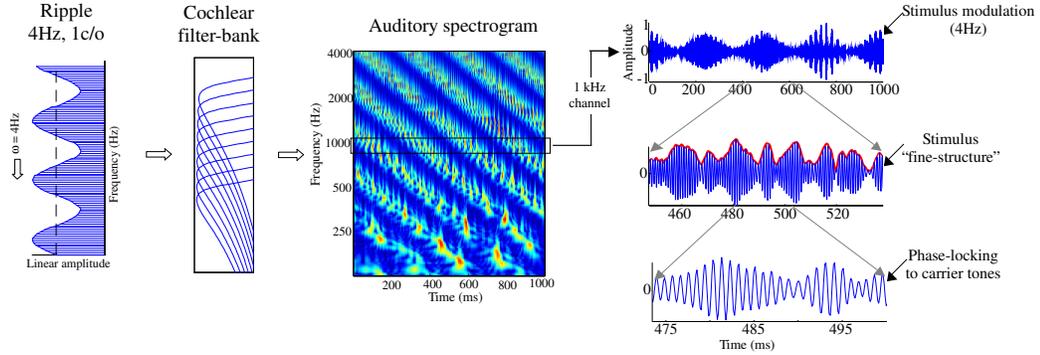


Figure 3.2: Schematic of the stimulus envelope and fine structure. *Left*, a ripple stimulus (4 Hz and 1 cycle/octave) is given as input to a cochlear filter-bank. *Middle*, the time waveforms of the filter outputs (auditory spectrogram) show an overall pattern of a 4 Hz drifting spectrogram, with detailed fast fluctuations. For display purposes, the output of each filter is half-wave-rectified to reveal better the fluctuation patterns in the spectrogram. *Right*, top, the output of the 1 kHz channel reveals the 4 Hz envelope modulating a faster carrier. Middle trace, View of the channel output at a higher magnification reveals a 1 kHz carrier with a rapidly fluctuating envelope or fine structure (red curve). The fine structure is caused by interactions between the tones that fall within the bandwidth of the 1 kHz filter. Bottom trace, A more detailed look of the modulated output of the 1 kHz filter.

byproduct of cochlear filtering is the creation of amplitude-modulated complex waveforms arising from the beating or interaction between the carrier tones that fall within the pass-band of the filters. These complex waveforms are called the *fine structure* of the stimulus [61]. They can be extracted by a Hilbert transform of the filter output [115], as shown by the red trace in Figure 3.2. The dynamic range (rate of fluctuation) of the fine-structure waveforms increases as the cochlear filter bandwidths become broader at higher frequencies. Note that the fine-structure waveforms depend solely on the carrier of the ripple (or TORC), and are independent of the global envelope. Because we constructed all our stimuli with identical carrier tones, their fine-structure waveforms are identical.

### 3.1.2 Cortical response properties

Using TORC stimuli, we examine properties of neuronal responses collected extracellularly in the primary auditory cortex (A1) of ferrets (*Mustela putorius*). The data analyzed in this work was collected in the Neural System Laboratory in the context of various studies [49, 60, 69, 70] from a total of 8 domestic ferrets. Three of these ferrets were anesthetized during the experiments (full procedural details in [132]). The remaining five ferrets were used for awake recordings. Among the 5 animals used for the awake experiments, three ferrets were awake but were not trained on any behavioral task, while the remaining two were trained to perform an acoustic detection task while the recording was in session [70]. Tungsten electrodes (5-7 M $\Omega$ ) were used to record single and multi-unit responses at different depths. To isolate single-unit responses, we employed both automatic [96] and manual off-line spike-sorting procedures.

The data-set was based on cortical recordings from a total of 918 single units (37% from anesthetized animals). The awake recordings were typically characterized by a more vigorous firing rate; but apart from this difference, our analysis and findings apply to both anesthetized and awake conditions, unless otherwise stated. Most units encountered in both anesthetized and awake recordings responded in a sustained fashion to the TORC stimuli as illustrated by the 1 sec response rasters for the two example neurons in Figure 3.3 (Each point in the raster plot corresponds to an action potential, or a spike). The responses exhibit simultaneously two patterns of phase-locking. First, they are phase-locked to the TORC envelopes, as evidenced by the changing raster display from one TORC to the other. Second, the spikes are also precisely locked to the fine temporal patterns, common to all TORCs, giving the appearance of vertically aligned episodes in the raster plots across two or more TORCs. This dual pattern of locking explains

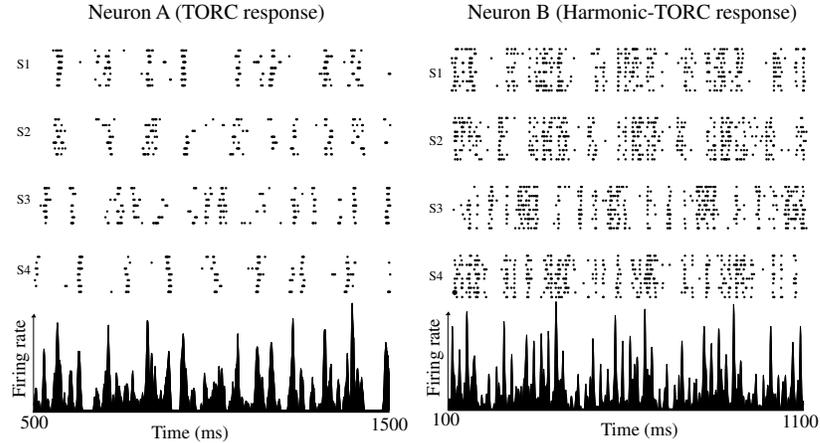


Figure 3.3: Rasters of A1 responses: Cortical responses of two single units in A1 of an anesthetized (left) and awake (right) animal. Each raster depicts repeated responses to four different TORC stimuli. The bottom panels depict the PSTHs computed by averaging the responses to repetitions of all TORC stimuli presented to that neuron. The precision of the time of occurrence of spikes can be judged by their vertical alignment. The PSTH contains frequent large peaks, indicating the occurrence of spikes at those instants in response to many of the TORCs. The TORCs in the right panel are composed of harmonically related tones (H-TORCs) with a fundamental frequency of 48 Hz. Therefore, the PSTH displays regular peaks locked to the 48 Hz fundamental.

the disappearance of the vertically aligned episodes in some TORCs. This can be seen more easily in the left panel of Figure 3.3 where the harmonic TORC elicits responses locked to the fundamental frequency (48 Hz) of the harmonic sequence that makes up the TORC carrier (and hence its fine-structure). The histogram in the bottom of Figure 3.3 (right panel) accumulates responses over all repetitions of all TORCs, and illustrates the regularly-spaced 48 Hz peaks due to the periodicity of the fine-structure. Note that while all raster spikes tend to occur at the regular 48 Hz intervals, they are completely missing in some TORCs, seemingly because the TORC envelope gates the occurrence of the spikes as we shall discuss later in more detail.

### 3.1.3 Temporal resolution of cortical information

The reliability of the neural responses observed in A1 is directly related to the stimulus-specific information carried by sensory neurons. This information could be encoded at any temporal resolution driven by very different temporal patterns in the input sound. To explore the temporal resolution of the neural code, we use a direct approach that relies on statistical properties of the neuronal responses themselves. This analysis involves only comparisons of spike trains with no reference to specific stimulus parameters.

#### Spike correlation methodology

The temporal resolution of neural responses is reflected in two major parameters: **(1)** how reproducible each spike train is (i.e., is the same spike count reproduced from one trial to the other), and **(2)** how accurate is the neural response at different trials (how much jitter exists in the reproduction of each spike?). To address these questions, we define an across-trial spike train correlation function from the neural responses of each neuron. Such function is described by the equation:

$$r(\tau) = \frac{2}{MN(MN - 1)} \sum_{i=1}^{MN} \sum_{j < i} r_{i,j}(\tau) \quad (3.2)$$

for  $M$  stimuli and  $N$  trials of each stimulus, and where  $r_{i,j}(\tau)$  is the cross-correlation between the  $i^{th}$  and  $j^{th}$  response traces. Effectively, this equation describes the correlation of the response to each TORC sound with responses to (about 10) repetitions of the same stimulus, as well as responses to all other stimuli (a total of 30 stimuli). While Equation 3.2 hints to intense computations (about 45,000 correlation operation for  $M=30$  and  $N=10$  on average), it can in fact be implemented in a greatly simplified way by taking advantage of the fact that each spike train is a binary signal (a sequence of 1's -spikes- and 0's -no

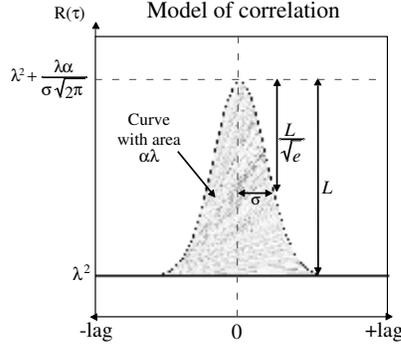


Figure 3.4: Model of the expected correlation function for a Poisson spiking neuron. The model is used to estimate spiking jitter ( $\sigma$ ), spike reproducibility ( $\alpha$ ), and average firing rate ( $\lambda$ ) of neuronal responses.

spikes-). In this study, the correlation function for each neuron was computed over a window of  $\pm 100\text{ms}$ , with a spike resolution (bin-size) of  $1\text{ms}$ .

Using the correlation function defined in Equation 3.2, we would like to extract parameters that reflect spike-timing jitter and spike reproducibility. To do so, we derive an analytic correlation function based on a model of Poisson correlations [119]. Assuming that each spike train is a realization of a Poisson-point process, its correlation function is given by:  $\hat{r}(\tau) = \lambda^2 + \lambda\delta(\tau)$  ([119], pg. 287). The parameter  $\lambda$  corresponds to the spike-rate of the process. By modelling spike jitter as a normal distribution with parameter  $\sigma$ , we can replace the delta function in  $(\hat{r})(\tau)$  by a Gaussian function. Additionally, we introduce a variable  $\alpha$  to control the area under the normal distribution.  $\alpha$ , which varies between 0 and 1, reflects the probability of spike reproducibility ( $1-\alpha$  corresponds to the probability of spike deletion). A probability  $\alpha$  equals 1 indicates a Gaussian distribution with a total area of 1, and thus perfect reproducibility of spikes. As  $\alpha$  approaches zero, the probability of spike reproducibility decreases, and thus, the peak of the correlation function is reduced. Therefore, the overall analytic model to fit our correlation function is depicted in Figure 3.4, and captured by the equation:

$$\hat{r}(\tau) = \lambda^2 + \frac{\alpha\lambda}{\sigma\sqrt{2\pi}}e^{-t^2/2\sigma^2} \quad (3.3)$$

## Spike correlation results

We fit the data function  $r(\tau)$  (eq. 3.2) with the Poisson-based model  $\hat{r}(\tau)$  (eq. 3.3), and extract the parameter triplet  $(\alpha, \sigma, \lambda)$  for each neuron. These parameters are direct correlates of the extent and accuracy of cortical phase-locking to the TORC fine-structure.

The range of values observed over our data set is illustrated in Figure 3.5. Over 63% of anesthetized and 77% of awake recordings exhibited precise locking of less than 10ms accuracy ( $\sigma \leq 10\text{ms}$ ). Note also that the awake population exhibited on average higher precision ( $\sigma_{mean}$  of 18.7 vs. 11.7 ms in the middle panels of Figure 3.5). The distribution of the spike reproducibility parameter ( $\alpha$ ) was biased towards 0 under all experimental conditions (Figure 3.5, left panels). This is partly due to the fact that envelopes of different TORC stimuli are uncorrelated; and hence spikes are suppressed (gated out) differently from one TORC response to another. Therefore, computing a correlation function across responses to all stimuli would exhibit a reduction of spike reproducibility. Finally, the spike rate in the awake population was expectedly [60] significantly higher than in the anesthetized ( $\lambda_{mean}$  of 18 vs. 9 as indicated in the left panels of Figure 3.5).

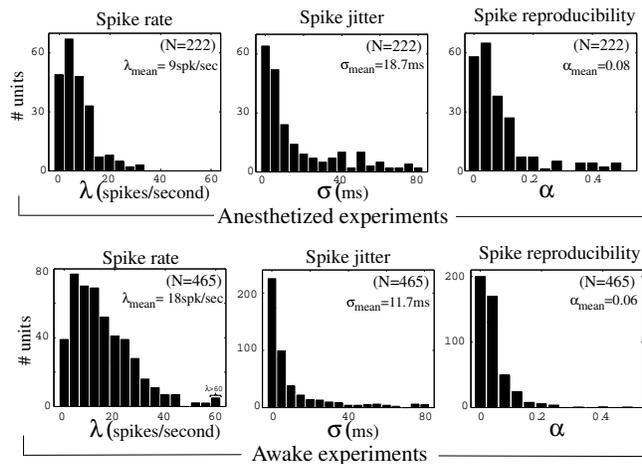


Figure 3.5: Distribution of correlation model parameters. The histograms are population distributions for average spike rate ( $\lambda$ ), spike jitter ( $\sigma$ ), and reproducibility ( $\alpha$ ), under anesthetized and awake conditions.

## 3.2 Modelling cortical responses

So far, we have shown that the primary auditory cortex *can* respond precisely not only to sound onsets, but also to rapid *sustained* stimuli with precision of the order of a few milliseconds. This precise spiking to the fine structure of long-duration stimuli appears to be contingent on the presence of relatively slow modulations of the stimulus spectrotemporal envelopes that can effectively excite the cortex. By using a specially designed acoustic stimulus (the TORC), we can model the system properties that give rise to these responses.

### 3.2.1 Defining a piecewise linear model

We start by outlining the methodological basis for modelling the dynamical system under study. The key strategy of our analysis is to approximate the nonlinear mapping between sensory inputs and cortical neural responses by a collection of *piecewise linear systems* [117]. Conventionally, STRFs (Spectrotemporal receptive fields) have been used to describe the *linear* relationship between the neural response of a unit and the dynamic spectrum (time and frequency-dependent energy) of a stimulus. Various studies using the STRF technique have tailored it to characterize the response sensitivity of neurons to the stimulus modulation profile (or spectrotemporal envelope) [49, 108, 129]. The choice of this approach was motivated by physiological evidence relating cortical neuronal responses to the modulation frequencies of a stimulus profile.

In this work, we wish to expand the STRF-based technique to explore the time constants of cortical neurons. We base our analysis on the theory of piecewise linear (PWL) systems. PWL is a system identification technique which approximates a nonlinear dynamical system via multiple linearizations at different operating points [98]. A classic

approach in PWL system identification is to choose *a priori* defined gridding of the input space to define linearization regions [19, 86]. Within each region in this grid, a linear approximation of the system is applied around a different operating point. Here, we follow the gridding theory introduced by Billings and Voon [19], where the input space is divided into rectangular sets with sides parallel to the coordinate axes. This approach fits in nicely with our analytic description of the ripple space in terms of its Fourier components (Figure 3.1). The linearization can then be performed on each sub-region of the ripple space (envelope and fine structure), as we show next.

### 3.2.2 Cortical functions with dual operating-point

The choice of the gridding  $\{\Lambda_E, \Lambda_F\}$  of the stimulus space comes as a natural choice reflecting both our stimulus construction methodology, as well as the definition of a PWL model. In order to capture details of cortical encoding of both the stimulus envelope as well as its fine structure, these two regions in ripple space are defined as the two operating points for linearization of the system functional  $\mathcal{F}[\cdot]$  (Equation 2.4). Within each stimulus partition, we employ standard linear techniques, namely the STRF approach through reverse correlation [57]. Mathematically, the STRF is defined by the equation:

$$r_{lin} = \int S(t, x) *_t STRF(t, x) dx \quad (3.4)$$

where the linear component of the firing rate  $r_{lin}(t)$  is described by a convolution in time ( $t$ ) and integration over logarithmic frequency ( $x$ ) of the spectrotemporal stimulus representation  $S(t, x)$  and the  $STRF(t, x)$ . Most STRF measurements are made by reverse-correlating (or convolving) the stimulus spectrogram with the responses of the cell:

$$STRF(t, x) = \int dt'' \int dx' \int dt' M^{-1}(t, x) S(t' - t'', x') r(t') \quad (3.5)$$

where  $M(t, x; t', x') = \int dt'' S(t'' - t, x) S(t' - t'', x')$  is the spectrotemporal autocorrelation of the stimulus.

In this study, we define two linear functionals describing the linear processing in the two partition  $\Lambda^E$  and  $\Lambda^F$ , and corresponding to two linear functions  $\text{STRF}^E$  and  $\text{STRF}^F$ . Such estimates try to map different stimulus regions, and thus they differ in the choice of the stimulus representation  $S(t, x)$ . The two STRFs derived in this case are:

1.  $\text{STRF}^E$  (for Envelope), which uses the stimulus profile or spectrotemporal envelope for reverse correlation (see Figure 3.6(A)). In this case, the stimulus autocorrelation function ( $M(t, x; t', x')$  in Equation 3.5) is straightforwardly defined in the Fourier domain; as explained in detail in [88]. Note that the superscript “E” is used to easily identify the STRF measured from the stimulus envelope.
2.  $\text{STRF}^F$  (for Fine structure), which captures solely the spectrotemporal patterns in the stimulus fine structure that selectively drive the neuron (independently of the stimulus envelope). In this case, we use fine-structure profiles (see Figure 3.6(B)) as a stimulus trigger for reverse correlation. These fine-structure profiles are obtained by averaging the complete profiles of all TORC stimuli. In this case, the stimulus autocorrelation is not trivial: it is approximately a periodic delta function in  $t - t'$ , whose period depends on  $x$  and  $x'$  (Figure 1 in [61]). Approximating it as an exact delta function (the standard autocorrelation) gives the correct  $\text{STRF}^F$  but with an occasional periodic artifact at high spectral frequencies.

These STRFs reveal the differential selectivity of cortical cells to these two sources of information in the acoustic stimulus. The relationship between  $\text{STRF}^E$  and  $\text{STRF}^F$  reflects the interaction between cortical slow and fast dynamics. To further explore this

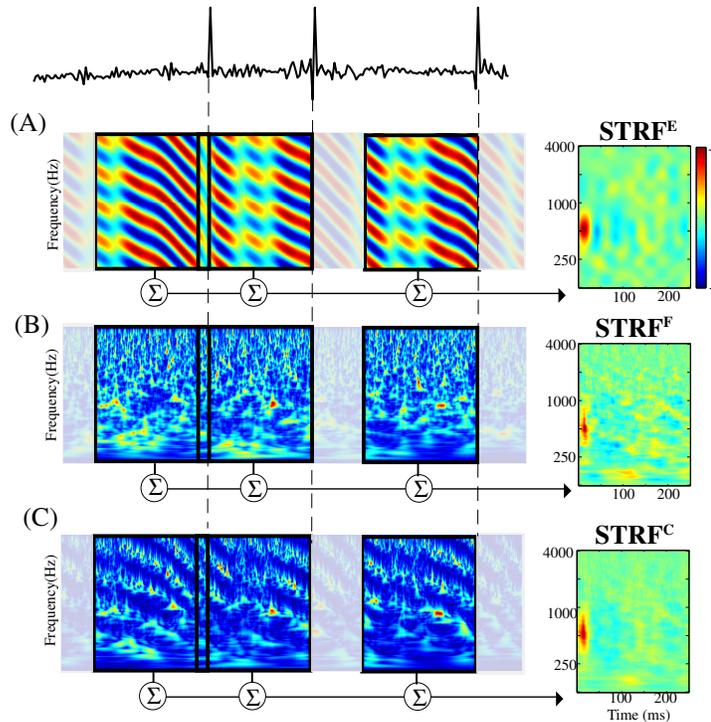


Figure 3.6: Schematic illustrating the use of reverse-correlation to derive STRFs. Top, Trace of a neuronal response. **A–C**, Three spectrotemporal representations of the TORC stimuli. In all three cases, the stimulus profiles preceding the occurrence of action potentials are averaged. Because the stimulus time evolves from left to right, the actual spike-triggered average is a stimulus spectrogram represented from -250 to 0 ms, where 0 ms corresponds to the actual occurrence time of the spike. The final STRF is a time-reversed average stimulus spectrogram, in which the time axis is flipped and the STRF can then be interpreted as the receptive field of the neuron.

point, we can also employ a complete linearization of the system. We hence use a stimulus representation  $S(t, x)$  that covers the entire stimulus space, and use it to derive what would be a first order linear approximation of the entire system. Such functional is termed  $\text{STRF}^C$  and uses a Complete spectrotemporal representation  $S(t, x)$  of the stimulus including both its envelope and fine-structure patterns (see Figure 3.6(C)). This spectrographic representation of the stimulus captures both its envelope dynamics, which are comparable with the envelope profile in Figure 3.6(A), as well as the stimulus fine temporal structure,

created by the interaction of the TORC carrier tones during the filtering process. In this case, the stimulus autocorrelation is a complex combination of the envelope autocorrelation and fine-structure autocorrelation (Figure 1 in [61]).

### Interplay of the two STRF operating points

One way to relate these three STRFs is to use an estimate of their *total power* defined as:

$$P = \frac{1}{TX} \int_0^T \int_0^X STRF(t, x) dt dx \quad (3.6)$$

By contrasting the total power  $P_T$  in the  $STRF^C$  over the entire ripple space, to the power  $P_E$  of the  $STRF^C$  in partition  $\Lambda^E$ , we can estimate the contribution of the envelope patterns to the total response represented by  $STRF^C$ . Such contribution is captured by the ratio  $\Delta P_E = P_E/P_T$ . The remainder remainder of the power ( $1 - \Delta P_E$ ) is ascribed to the faster modulations of the  $STRF^F$ . Clearly, this estimate assumes the two ranges of  $STRF^E$  and  $STRF^F$  modulations are mutually exclusive, and hence ignores the relatively small contribution of the fine structure to the slow modulations in the  $STRF^E$  range. Nevertheless, this approximation is adequate for our purposes, and we will assume that the  $STRF^C$  is composed of a linearly weighted sum of  $STRF^E$  and  $STRF^F$  in the proportions of their power estimates ( $\Delta P_E$  and  $1 - \Delta P_E$ ) (see Figure 3.8):

$$\begin{aligned} STRF_{predicted}^C &= \Delta P_E \cdot STRF^E + (1 - \Delta P_E) \cdot STRF^F \\ \rho &= \langle STRF^C, STRF_{predicted}^C \rangle \end{aligned} \quad (3.7)$$

A robust linear behavior of the STRFs would result in  $STRF^C$  predicted being similar to the measured  $STRF^C$ . We use a correlation coefficient [119] as a measure of similarity between the original  $STRF^C$  and  $STRF_{predicted}^C$ . The correlation coefficient  $\rho$  takes values

between +1 and -1, with +1 indicating a perfect match between the two STRF measures. The correlation puts our assumption of linearity to the test, and shows how much of the STRF<sup>C</sup> power is captured by a linear sum of the powers in STRF<sup>F</sup> and STRF<sup>E</sup>.

### 3.2.3 Cortical receptive fields

#### STRF examples

STRFs in A1 exhibit a wide range of shapes and forms, reflecting the immense variety by which A1 units process and integrate various stimulus features along the spectrotemporal dimensions. Figure 3.7 displays several examples of receptive-field triplets (STRF<sup>E</sup>, STRF<sup>F</sup>, and STRF<sup>C</sup>) obtained for different neurons. Generally, the STRF<sup>C</sup> displays features that are prominent for both the STRF<sup>E</sup> and STRF<sup>F</sup>, depending on the contribution of each to the total power. The value in the lower right corner of each STRF<sup>E</sup> and STRF<sup>F</sup> indicates its estimated contribution to the overall STRF<sup>C</sup> of that neuron, as captured by the values of  $\Delta P_E$  and  $1 - \Delta P_E$ . Apart from the center frequency of the STRFs, we found no obvious relationship between the shapes of the STRF<sup>E</sup> and STRF<sup>F</sup>.

The left column of Figure 3.7 shows a selection of neurons characterized by the large contribution of envelope features to their overall STRF<sup>C</sup>, and hence the close similarity between their STRF<sup>C</sup> and the STRF<sup>E</sup>. Figure 3.7(A) is a classic example of a broadband offset unit, which preferentially responds to the offset of a stimulus over a wide frequency range. The broad tuning of this unit explains the weak contribution of the STRF<sup>F</sup> because integrating from a large number of cochlear channels results in a complex waveform that is weakly correlated with any particular channel. Figure 3.7(B) illustrates an example of a spectrotemporally rich STRF with a similar STRF<sup>C</sup>. The STRF<sup>F</sup> here is rather simple, completely lacking the inhibitory fields of the STRF<sup>E</sup>. Finally, Figure 3.7(C) depicts an

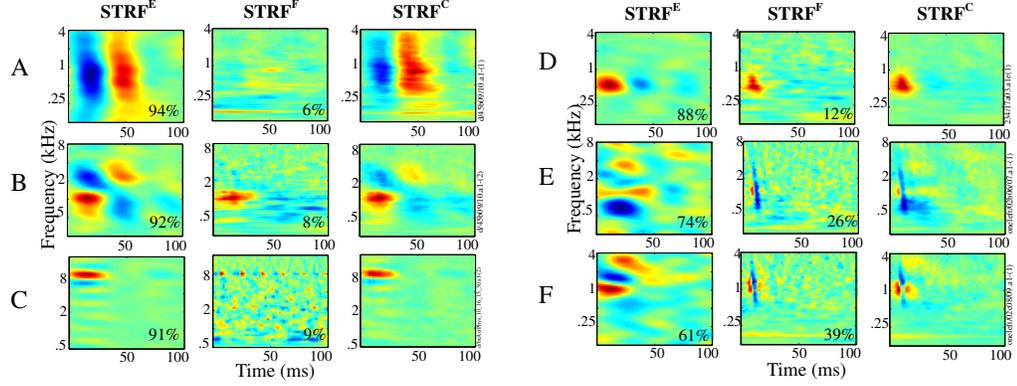


Figure 3.7: Examples of STRFs, with gradually increasing contributions of fine structure (from 5 to 40%). Each STRF triplet in each row corresponds to the  $\text{STRF}^E$ ,  $\text{STRF}^F$ , and  $\text{STRF}^C$  derived for one neuron. An estimate of the contributions of  $\text{STRF}^E$  and  $\text{STRF}^F$  to the total power of the  $\text{STRF}^C$  is indicated in each panel. Each triplet is individually scaled to span the full range of colors in the color map. The fine-structure characteristics of the cells shown in this figure are as follows: **(A)**  $\lambda$ : 3.9 spikes/sec,  $\sigma$ : 10 ms,  $\alpha$ : 0.16,  $\Delta P_E$ : 94%; **(B)**  $\lambda$ : 10.7 spikes/sec,  $\sigma$ : 10 ms,  $\alpha$ : 0.07,  $\Delta P_E$ : 92%; **(C)**  $\lambda$ : 27.5 spikes/sec,  $\sigma$ : 1 ms,  $\alpha$ : 0.01,  $\Delta P_E$ : 91%; **(D)**  $\lambda$ : 8.84 spikes/sec,  $\sigma$ : 4 ms,  $\alpha$ : 0.3,  $\Delta P_E$ : 88%; **(E)**  $\lambda$ : 10.3 spikes/sec,  $\sigma$ : 1 ms,  $\alpha$ : 0.05,  $\Delta P_E$ : 74%; **(F)**  $\lambda$ : 12.8 spikes/sec,  $\sigma$ : 1 ms,  $\alpha$ : 0.03,  $\Delta P_E$ : 61%.

example of a high-frequency cell, with a simple excitatory field at about 8 kHz. The  $\text{STRF}^E$  shares very similar features with the  $\text{STRF}^F$ , with the exception of the much faster temporal dynamics in the latter. The periodic structure of the  $\text{STRF}^F$  is a result of the fact that the TORC carrier tones near 8 kHz are approximately equally separated within the narrow bandwidth of the  $\text{STRF}^E$ , hence creating a pseudo-periodic carrier waveform whose autocorrelation is also periodic.

The units depicted in the right column of Figure 3.7 are highly influenced by the fine-structure features, because they all exhibit a relatively high contribution of the  $\text{STRF}^F$  to the overall response (i.e., lower  $\Delta P_E$  values). Figure 3.7(D) illustrates an example of change in temporal dynamics in response to the stimulus fine structure. The  $\text{STRF}^F$  of this unit shares the excitatory field with the envelope-based  $\text{STRF}^E$  at about 500 Hz,

but its temporal extent is much more narrow, and lacks any inhibitory surround. The example in Figure 3.7(E) illustrates a unit with very rapid temporal selectivity for the  $\text{STRF}^F$ . Finally, Figure 3.7(F) is a striking example of the independence of the fast and slow temporal features in cortical STRFs. The  $\text{STRF}^E$  of this unit exhibits two excitatory fields surrounding an inhibitory region near the best frequency (BF) at 1 kHz. However, its corresponding  $\text{STRF}^F$  indicates a specific selectivity to particularly fast oscillatory temporal patterns (at about 150 – 200 Hz). This selectivity is reflected in the consecutive excitatory and inhibitory fields in the  $\text{STRF}^F$  (about 2 kHz). In turn, this pattern strongly dominates the  $\text{STRF}^C$ . Note, however, that despite the similarity of the  $\text{STRF}^F$  shapes of the different neurons in Figure 3.7(E,F), their corresponding  $\text{STRF}^E$ s are very different, demonstrating again the independence of these two sources of information processing.

### Relating the two functions

We examined the response fields that emerge when taking into account neuronal responses to the envelope alone ( $\text{STRF}^E$ ), fine structure alone ( $\text{STRF}^F$ ), or the combined features ( $\text{STRF}^C$ ). These STRFs (Figure 3.8(A)) reveal the differential spectrotemporal selectivity that cortical cells exhibit to these two sources of information in the acoustic stimulus, as we will discuss below. Figure 3.8(A) illustrates an example of an STRF triplet derived from the responses of one neuron. To demonstrate the relationship between these three STRF descriptions, we computed the proportion of the power contributed to the  $\text{STRF}^C$  by the envelope and fine-structure sources. The two-dimensional Fourier transforms of all three STRFs of this neuron are shown in Figure 3.8(A). The black box delimits the energy region spanned by the TORC envelopes. By construction, the envelope-based  $\text{STRF}^E$  is defined only over the range  $\Lambda^E$ , and thus contains no energy outside this area. In contrast, the

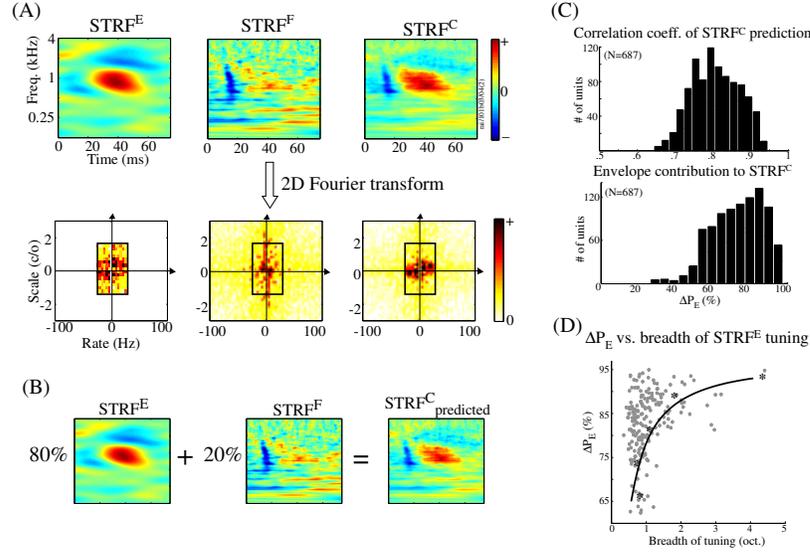


Figure 3.8: Example of an STRF triplet of a neuron and its significance. **(A)** The three STRFs (STRF<sup>E</sup>, STRF<sup>F</sup>, and STRF<sup>C</sup>) of a neuron, depicted both in the time-frequency and Fourier domains. The black box delimiting a subregion of the Fourier domain marks the range of spectrotemporal modulations spanned by the TORC stimuli (area  $\Lambda^E$ ). The fine-structure characteristics of the cell shown in this figure are as follows:  $\lambda$ : 17.8 spikes/sec,  $\sigma$ : 3.5 ms,  $\alpha$ : 0.02,  $\Delta P_E$ : 80%. **(B)** Estimating the contributions of the envelope and fine structure to reconstruct the STRF<sup>C</sup>. **(C)** Top, Distribution of the correlation coefficient relating the STRF<sup>C</sup> to its prediction using  $\Delta P_E$  (Equation 3.7). Bottom, Distribution of values of  $\Delta P_E$  observed in our data set. **(D)** Scatter plot of  $\Delta P_E$  variations as a function of breadth of tuning of the STRF<sup>E</sup>. The solid curve is the best exponential fit to the means of the data within  $\pm 3.5\%$  around each  $\Delta P_E$ . The mean points are shown as asterisks.

fine-structure spectrograms include both coarse and fine temporal and spectral patterns, and thus the energy content of the STRF<sup>F</sup> spreads over a wider range of spectral and temporal modulations.  $\Delta P_E$  is computed from this representation as the ratio of the power within the box to the total power. For this neuron,  $\Delta P_E = 0.8$ , which results in a predicted STRF<sup>C</sup> (Figure 3.8(B)) that strongly resembles the measured STRF<sup>C</sup> (Figure 3.8(A)). Such resemblance has been observed for most units that exhibited highly precise responses, and for which we successfully derived an STRF<sup>F</sup> (about 70% of units). This

result supports the notion that the linear component of responses in A1 is very robust and strongly captured by the STRF descriptors. Figure 3.8(C) shows the distribution of correlation coefficients between the STRF<sup>C</sup> and the linearly predicted complete STRF derived from all units. The distribution indicates a high degree of correlation, with mean coefficient of +0.83 confirming the high degree of linearity in A1 responses, and thus suggesting an *independence* of the expression of envelope and fine structure in cortical responses. The range of values of  $\Delta P_E$  found in all units is shown in Figure 3.8(C). This distribution is biased toward higher values of  $\Delta P_E$ , indicating that the majority of units are driven primarily by their responses to time-varying spectral envelopes. This result is consistent with the accepted notion that A1 is particularly sensitive to slowly varying modulation patterns [49, 56]. Nevertheless, over half of all cells exhibit a significant contribution ( $\leq 25\%$ ) to their STRF<sup>C</sup> from the fine-structure modulations, indicating that regular regular envelope-based STRF measurements are insufficient to capture all relevant spectrotemporal features of their response fields.

### **Responses to harmonic complexes**

In 117 units, we recorded cortical responses using harmonic TORCs (also called H-TORCs) as well as regular TORCs. Because of their regular structure, harmonic TORCs evoke periodic phase-locked responses that reflect the fundamental frequency of the stimulus carrier. Therefore, it is particularly easy to discern visually and computationally the degree of neuronal locking to fine temporal structure of the stimulus. For instance, one simple indicator of locking to the fine structure is the prominence of the Fourier coefficient at the fundamental frequency, computed from the Fourier transform of the average PSTH of all harmonic TORC responses. Recall that taking the average PSTH eliminates locking

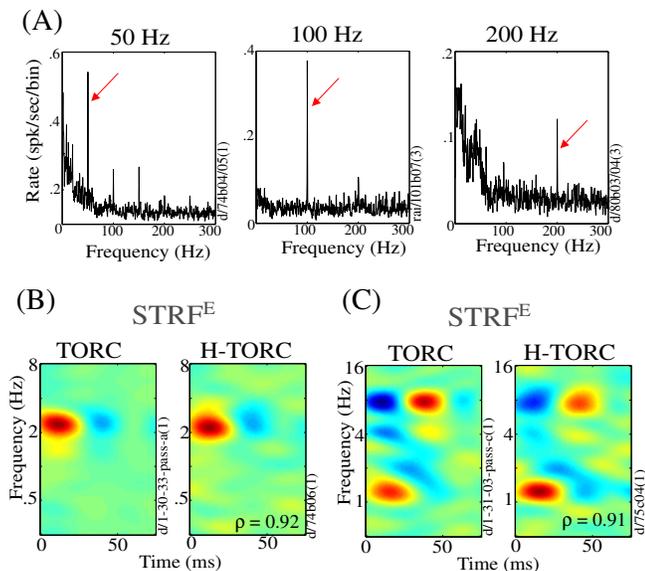


Figure 3.9: Harmonic TORC responses. **(A)** Fourier transform of PSTHs of three neurons. Each neuron was tested with a different harmonic series, as indicated by the fundamental frequencies marked by the red arrows. All three neurons exhibit noticeably salient peaks at the fundamental component, and some of the upper harmonics. **(B,C)**  $\text{STRF}^E$ s estimated using regular TORCs (left) and H-TORCs (right) for different neurons. The  $\text{STRF}^E$  pairs are very similar, with correlation coefficients of  $+0.92$  and  $+0.91$ . The fine-structure characteristics of the cells shown in **(B)** and **(C)** are as follows: **(B)**  $\lambda$ : 45.6 spikes/sec,  $\sigma$ : 3 ms,  $\alpha$ : 0.2,  $\Delta P_E$ : 93%; **(C)**  $\lambda$ : 20.3 spikes/sec,  $\sigma$ : 8.5 ms,  $\alpha$ : 0.01,  $\Delta P_E$ : 90%.

to the TORC envelopes, because these are uncorrelated across different TORCs. Figure 3.9(A) shows examples of this spectral analysis from 3 units. The red arrow points to the peak corresponding to the spectral component (Fourier coefficient) at the fundamental frequency used in the stimulus. All 3 units have strong locking to the harmonic fundamental frequency, up to 200 Hz. Of the 117 neurons tested, approximately half displayed noticeable locking to their harmonic fundamental frequency (over the range 25 – 200 Hz). This finding is remarkable for A1 units that are generally incapable of following sustained temporal modulations beyond 20 Hz.

We exploited the fact that the harmonic TORCs and regular TORCs stimuli share

the same envelope structure to extract and compare their envelope-based STRF<sup>E</sup>s. The goal of such a comparison is to determine whether cortical processing of the envelope is affected by the exact nature of the fine structure of the stimulus. Figure 3.9(B,C) demonstrates that the STRF<sup>E</sup>s derived from either type of TORC are very similar for both units. Such similarity has been observed for all units for which we recorded a full set of TORCs and H-TORCs to derive a pair of STRF<sup>E</sup>s using both types of stimuli. In all these cases, comparing the STRF<sup>E</sup> obtained from TORCs and harmonic TORCs indicates a high degree of correlation between the two, with all correlation coefficients greater than +0.5, and mean +0.75. This finding strongly supports the notion that the cortical representation and processing of the envelope and the carrier do not seem to influence each other substantially.

We will next explore the hypothesis that the envelope responses play a modulatory role for the expression of the fast fine-structure responses. Finally, note that it is not possible to obtain an STRF<sup>F</sup> from H-TORC responses because all fine-structure patterns are at the same fundamental frequency.

### **Prediction of A1 responses**

To illustrate directly the contribution of each of the STRF<sup>E</sup>, STRF<sup>C</sup>, and STRF<sup>F</sup> to the description of the unit responses, we compared the actual responses to the TORC stimuli to those predicted using the STRFs. This is a common approach used previously to validate the linearity assumption underlying the definition and computation of the STRF [49, 90, 138]. We expected that the STRF<sup>E</sup> would predict a smoothed version of the PSTH of TORC responses, whereas the STRF<sup>C</sup> would predict a more detailed waveform that includes the fine structure. Figure 3.10 illustrates plots of the response of a cortical unit

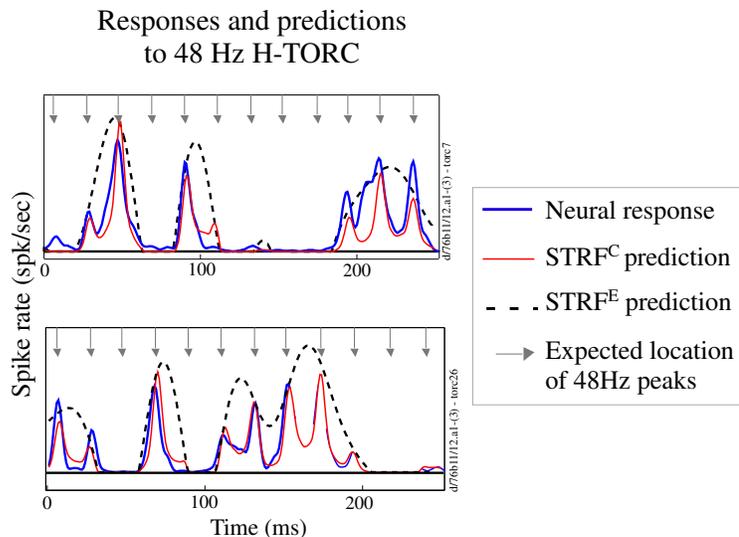


Figure 3.10: Comparison of actual and predicted responses to 48 Hz H-TORCs. Each plot illustrates a 250ms period histogram of the responses. Responses (blue) and predictions (red) demonstrate the gating of the fine-structure peaks (arrows) by the responses to the envelope (dashed line).

to H-TORCs along with the prediction of this response using the  $\text{STRF}^E$  and  $\text{STRF}^C$ . The arrows mark the anticipated locations of the fine-structure peaks because the carrier tones are multiples of a 48 Hz fundamental. As expected, the predictions demonstrate that the envelope waveform effectively gates or modulates the expression of the fine-structure peaks. Thus, when the predicted response to the envelope is small, the fine structure diminishes; when the response to the envelope is large, the peaks are well expressed in the PSTH.

### 3.3 Neural mechanisms

Why do dynamics of cortical responses differ from those observed in the thalamic inputs? Specifically, why do repetitive stimuli fail to elicit synchronized responses in A1 much beyond 20 Hz, a decade lower than typically found in the MGB? Such a significant slowdown is apparently not caused by simple global low-pass filtering of thalamic inputs

because cortical cells are transiently still able to encode faithfully the rapid fine structure of the stimuli. Two potential mechanisms are examined here, both known to be operative at the thalamocortical synapses and input layers of A1.

### 3.3.1 Synaptic dynamics

The first mechanism is the depressive character of the excitatory thalamocortical synapses. When subjected to continuous or rapid stimulation, these synapses become temporarily depressed/weakened as the supply of transmitter is exhausted. If the stimulus (thalamic input) is transiently turned off or reduced, the synapse can recover its strength in time for the next input. The potential rate at which the recipient cortical cell can respond to its fluctuating thalamic input depends critically on the dynamics of this recovery phase.

Computational models of synaptic dynamics are readily available in the literature [26, 29, 34, 139] and used to simulate responses in many cortical modalities including the auditory and visual cortices. They model well known experimental and theoretical findings that cortical responses phase-lock well up to 15–20 Hz and are generally incapable of following much more rapid sustained periodic stimuli [56, 78, 90, 111, 130]. For instance, model responses diminish in amplitude gradually as the input pulse rate increases beyond 15 Hz. At lower rates ( $<2$  Hz), the onset response to each input pulse becomes highly accentuated, and simultaneously, the response to the body of the pulse becomes relatively suppressed (upper panel of Figure 3.11(A)). Nonetheless, responses of depressing synapses maintain the fast carrier of the stimulation, even as the overall response level drops. Using a slowly modulated stimulus, the response -while following the slow envelope dynamics in the input-, occurs with a pattern dictated by the fast carrier. In fact, the carrier patterns are particularly prominent at the onsets of the modulation pulses, and hence any spikes

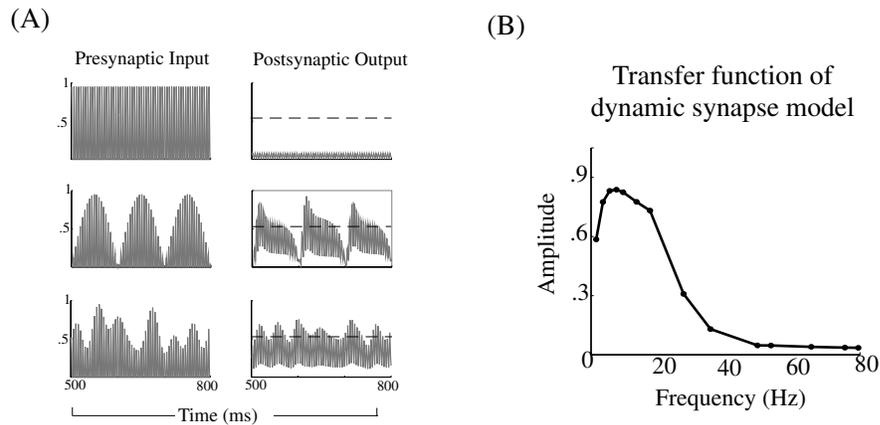


Figure 3.11: Modelling dynamic synapse mechanisms. **(A)** Dynamic synapse model input (presynaptic stimulus) and output (postsynaptic responses) to various modulations of a 200 Hz click train. Each row corresponds to the model responses to an unmodulated 200 Hz click train, 5-Hz-modulated click train, and TORC-modulated click train, respectively. The dashed line across the postsynaptic output panels represents the firing threshold. The simulation results of this figure are based on a model by Tosdyks *et al.* [139]. **(B)** Transfer function of the single dynamic synapse model in response to click-train stimuli.

that might be initiated by these onsets would likely reflect the timing of these peaks (Figure 3.11(A), middle and lower panels).

Such a scheme amounts to a net input-output transfer function with a band-passed shaped (Figure 3.11(B)), even though the dynamics of the model are inherently nonlinear. These nonlinear mechanisms guarantee the co-existence of slow and fast response patterns. Thus, when the (intracellular) response to the envelope is high, the fine structure associated with it rises and hence may exceed the spiking threshold causing precisely phase-locked action potentials to occur. Therefore, the ability of the model to respond to the fine structure is contingent on its ability to respond to the slow envelope, as they are the ones who gate the expression of the fine-structure peaks by allowing the synapses to “recover”.

### 3.3.2 Cortical circuitry

Depressing thalamocortical synapses are only one part of a complex cortical circuitry that involves co-activated excitatory and inhibitory influences that impinge on thalamo-recipient cortical cells. In fact, it has been postulated that such interactions, coupled with slow NMDA synapses, are themselves responsible for endowing the cortex with its characteristic dynamics and temporal tuning [82, 92, 106]. The question therefore arises as to whether a simple model circuit of an excitatory thalamocortical input and a concurrent slower, intra-cortical, feedforward inhibitory input could give rise to the type of dynamics observed in the cortex.

We know from physiological evidence that strong, feedforward, slightly delayed, and longer-lasting inhibition arrives after the onset of a persistent excitatory input. This inhibition reduces or suppresses the response, thus giving rise to the commonly seen phasic response at the onset of a stimulus. By slowly modulating the input strength ( $<20$  Hz), one can alter the relative phase of the inhibition and excitation, and hence reduce the mutual cancellation and increase the response. As the modulation rate is speeded up, it induces sustained inhibition that attenuates the response again (Figure 3.12(A)).

Clearly, both synaptic depression and cortical circuitry can both individually act as a plausible neural mechanism that could explain the temporal paradox observed in A1. These two mechanisms do not have to be mutually exclusive, and can co-exist together. Regardless of whether they act individually or together, they give rise to a net transfer function with a tuning around 5–15Hz (Figure 3.12(B)). The function has two main features: **(1)** a band-pass shape, which captures the tuning observed in cortical responses, as well as the low responsiveness to flat un-modulated stimuli; and **(2)** while tuned to a particular modulation range, it extends to much faster dynamics, explaining the presence

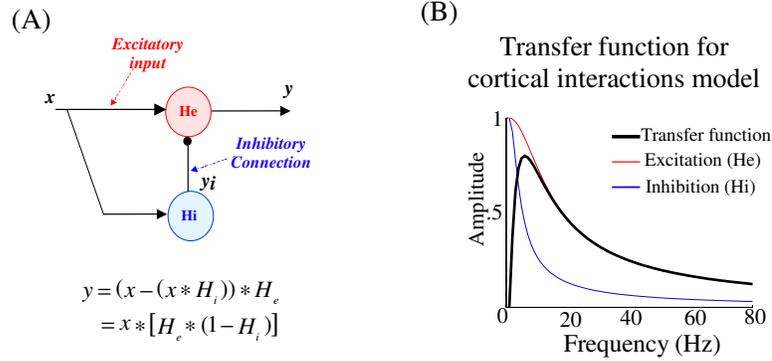


Figure 3.12: Modelling mechanisms of cortical interactions. **(A)** Schematic of cortical interactions between fast excitatory input (possibly thalamic input), and slow cortico-cortical inhibition. **(B)** Transfer function of the excitatory/ inhibitory cortical circuit with static weights, as well as the individual excitatory and inhibitory components ( $H_e$  and  $H_i$ ).

of precise timing in sustained cortical responses.

### 3.3.3 Emergence of cortical STRFs

Although synaptic depression and feedforward inhibition have been implicated and modelled to varying degrees at several sites along pre-cortical auditory pathways [105, 114], they are ubiquitous in all sensory cortices. This may explain the significant (order of magnitude) mismatch between the dynamics of the thalamus (medial and lateral geniculate nucleus) and cortex. It is therefore quite likely that the specialized processing of the spectrotemporal envelopes, as parameterized by the (envelope-based)  $\text{STRF}^E$ , is an emergent property exclusive to the cortex. In this new light, the  $\text{STRF}^E$  and  $\text{STRF}^F$  can be viewed as representing two distinct sources of information processing. The  $\text{STRF}^E$  reflects the explicit cortical extraction and processing of the stimulus spectrotemporal envelope and the information it conveys. In contrast, the precise spiking (phase-locked to the input fine structure) represents temporal dynamics inherited from pre-cortical stages [52, 94]; thus,

the STRF<sup>F</sup> measured from these precise firings provides a window to the spectrotemporal nature of the thalamic inputs to the cortex. The different origins of the STRF<sup>E</sup> and STRF<sup>F</sup> explain the apparent independence between their shapes, as shown in Figure 3.7.

Precise, rapid, and sustained spiking is however quite infrequently encountered in auditory cortical cells, though apparently common among many A1 neurons. Uncovering fast sustained dynamics in cortical responses requires stimuli that combine both a fine structure as well as a slowly modulated spectrotemporal envelope. In the absence of a fine structure, spikes phase-lock to the relatively slow envelopes of the inputs (2–20 Hz), and hence do not appear precisely timed except at sparsely spaced instants at which the envelope changes rapidly such as at stimulus onset. Similarly, stimuli with rapid fine structure but without spectrotemporal modulations (such as a sustained pure or complex tone, noise, or a fast click train) usually fail to elicit substantial response during their sustained portions, presumably because of adaptation, synaptic depression, or inhibitory influences [67, 71, 136]. Therefore, in a sense, the slowly modulated envelopes of acoustic stimuli gate temporally precise and sustained cortical responses. When the stimulus envelope is such that the “gate” is open, cortical cells can precisely phase-lock to the stimulus fine structure up to relatively high rates (>200 Hz). When the gate is closed in the absence of slow modulations, responses soon cease.

### 3.4 Functional significance

What is the functional significance and auditory perceptual correlates of precise cortical responses? Cortical cells respond well to change, manifested as modulated envelopes of carrier signals. The fine structure plays the important role of carrying these envelopes up to the cortex, where they are extracted and analyzed. Therefore, it is possible that the

precise spiking in the cortex reflects a pre-cortical carrier (fine structure), and that it has certain perceptual correlates such as: (1) the detection of rapid transient events in an otherwise slowly modulated signals such as speech [142], (2) the perception of “repetition” or “residue” pitch of  $< 400$  Hz [45, 133], and (3) the “roughness” or “texture” of the acoustic signal (e.g., the continuum between whispered and a pure voiced quality corresponding to the range from random to periodic fine structure [18, 104, 135]).

Additionally, we ascribed to synaptic depression and cortical circuitry the key innovation of the cortex: the creation of STRF<sup>E</sup>s to analyze and represent the spectrotemporally modulated envelopes of acoustic signals. These slow modulations are the main carrier of information in speech and music. In speech, they reflect movements and shape of the vocal tract, and consequently the sequence of syllabic segments in the speech stream. In music, slow modulations reflect the dynamics of bowing and fingering, the timbre of the instruments, and the rhythm and succession of notes. Analogously, spatiotemporal modulations in visual images are correlates of changing scenes and moving objects.

Overall, one may conjecture that a key role of cortical processing is in fact analyzing and representing the spectrotemporally modulated envelope of acoustic signals. These profiles are indeed the information bearing component of sounds. As we shall see in the next chapter, a direct correspondence between sound modulation and intelligibility of speech can be established based solely on the modulation content of the signal. In the following chapter, we explore the fact that the slow dynamics of cortical processing are commensurate with time constants associated with auditory streaming, while fast temporal events are essential for capturing the roughness or texture of acoustic signals and detection of transient acoustic events. The interplay between these temporal dynamics is in fact at the basis of what we presume is a cortical role in auditory scene analysis.

## Chapter 4

# Speech Intelligibility

Modulation spectra are the main carrier of information in speech and music [55], even though not sufficient for sound quality and music appreciation. They are so critical in perceiving sound that stimulations for cochlear implants preserve *only* the envelope attributes of sounds. We establish the relationship between sound modulations and intelligibility of sounds using a computational model of the auditory system. Intelligibility is defined by the ISO 9921 standard as “*a measure of effectiveness of understanding speech*” [3]. It is a critical measure for a wide range of applications such as transmitters design, room acoustics, hearing aids characterization, etc. However, it is not a physical quantity like loudness or voltage, making it an abstract percept that would require an objective metric for evaluating it. In this study, we demonstrate that mapping sound into a spectrotemporal modulation space can accurately predict the intelligibility level of a signal under any

noise condition. We establish the validity of the algorithm by performing a psychoacoustic study to correlate human estimates with model predictions.

We start by presenting an algorithm for a new intelligibility metric (the STMI), and validate its applicability to conditions of white noise and reverberation, by comparing its estimates to those of already existing intelligibility measures as well as psychoacoustic human scores. Next, we demonstrate the STMI performance for severe noise conditions under which existing measures fail, hence establishing the superiority of the suggested STMI metric. The use of a biologically inspired metric is also valuable in elucidating the role of the cortical representation of spectrotemporal modulation in the intelligibility and perception of sounds. The work presented in this chapter has been published in [59].

## 4.1 Measures of intelligibility

The articulation index (AI) and speech transmission index (STI) are the most widely available predictors of speech intelligibility up till now [6, 83, 93], and have proven to be extremely valuable in a wide range of applications ranging from architectural designs to vocoder characterization [20, 83, 84, 134]. The AI is an intelligibility metric that uses an estimate of the signal-to-noise ratio in various bands of speech. It has originally been developed in the late 1940's, and was later established by the American National Standards Institute as the speech intelligibility index (SII) ANSI-S3.5 [6]. Its main contribution is demonstrating the importance of different frequency bands in the speech spectrum. However, its applicability to various noise types is very limited, since it fails to effectively account for reverberation or any non-stationary distortion, and hence is rarely applicable in real-life situations. The STI, on the other hand, is a more realistic intelligibility measure. It has been developed by Steeneken and Houtgast in the 1970's at the well-known *TNO*

*Human Factors Laboratory* in the Netherlands. The STI's main improvement over the AI is the attempt to include distortions in the time domain, by evaluating the modulation-reduction in speech-like sounds. The STI uses modulated noise sounds as test signals to measure the reduction in modulation depth across a test channel, and maps it to loss of intelligibility. The Measurement of STI is defined by the International Electrotechnical Commission standard IEC 60268-16 [2].

In an effort to understand the underlying biological mechanisms that render such intelligibility measures meaningful, and how noise in general compromises the perception of speech and other complex dynamic signals, we want to employ a computational model directly relating our knowledge of the biology of hearing to the percepts of speech intelligibility. We use the model described in chapter 2, based on the neural evidence presented in chapter 3 as well as psychoacoustical measurements of human spectrotemporal modulation transfer functions (MTF) [31, 53]. Based on the premise that faithful representation of these modulations is critical for perception of sound [43, 53], we derive a novel intelligibility index, the *SpectroTemporal Modulation Index* (STMI), which quantifies the degradation in the encoding of spectral and temporal modulations due to any noise condition [59]. The STI, as we shall discuss later, is well suited to describe the effects of spectrotemporal distortions that are *separable* along the spectral and temporal dimensions, e.g. static noise (purely spectral) or reverberation (mostly temporal). The STMI, on the other hand, is a major elaboration on the STI in that it incorporates explicitly the *joint* spectrotemporal dimensions of the speech signal. As such, we expect it to be consistent with the STI in its estimates of speech intelligibility in noise and reverberations, but also be applicable to cases of joint (or inseparable) spectrotemporal distortions that are unsuitable for STI measurements, as with certain kinds of severe nonlinear distortions of speech or phase-

jitter and amplitude clipping in communication channels. Finally, like the STI, the STMI effectively applies specific weighting functions on the signal spectrum and its modulations; these assumptions arise naturally from the properties of the auditory model and hence can now be ascribed a biological interpretation.

## 4.2 Intelligibility of communication channels

This section introduces the use of STMI for intelligibility in communication systems. We call this variant  $STMI^R$  (R, for ripples) as shall become clear next. We first define what the  $STMI^R$  is, and present an algorithm for its computation. Then, we present  $STMI^R$  estimates for intelligibility under conditions of white noise and reverberation, and compare them to STI predictions.

### 4.2.1 $STMI^R$ procedure

Conceptually, the STMI is a measure of speech integrity as viewed by a model of the auditory system. In order to characterize intelligibility of a communication system (e.g., a recording or transmission channel, a room, or a vocoder), we use the auditory model to estimate the change in the spectrotemporal modulations that a test signal undergoes. We use ripples (see Equation 3.1) as test signals, and quantify the difference in the spectrotemporal modulation content of clean and noise-contaminated ripples, as analyzed through the auditory model (Figure 4.1). Ripples are ideal signals to characterize the Modulation Transfer Function (MTF) of the model, both with and without noise. The MTF is defined as the collection of responses of the auditory model filters to ripples at all temporal and spectral modulations. It determines the model's sensitivity to spectral and temporal modulations, and reflects how well these input modulations are faithfully transmitted through

the model.

Based on a clean reference and a noisy target channel (representing the transmission system under study), we quantify the difference between clean and noisy transfer functions, and map this distance metric to a direct measure of channel fidelity in transmitting ripple, and hence speech intelligibility in the system. We choose a Euclidian distance between the clean transfer function (MTF) and the noisy one (MTF<sup>\*</sup>). The exact procedure for STMI<sup>R</sup> computation is described in algorithm 1, with a reference to the auditory spectrogram and cortical filter-bank analysis presented in chapter 2. The algorithm starts by measuring the transfer function of the auditory model sensitivity to all spectrally and temporally modulated ripples in the range 2-32 (Hz) and 0.25-8 (cycles/octave). The same procedure is used to measure the transfer function of the noisy channel (MTF<sup>\*</sup>). The STMI<sup>R</sup> is then taken to be the distance between MTF and MTF<sup>\*</sup>.

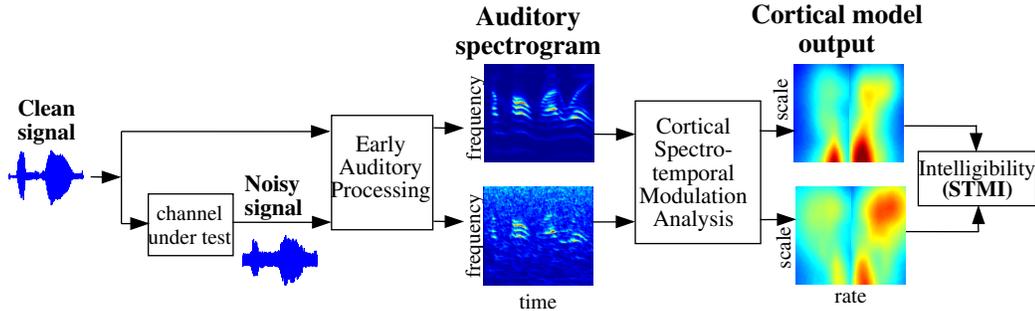


Figure 4.1: Schematic of STMI computation. The clean and noisy speech signals are given as inputs to the auditory model. Their outputs are normalized by the base signals as explained in the text. The right panel shows the cortical output of both clean and noisy inputs. These cortical patterns are then used to compute the template-based STMI.

---

♣ **Algorithm 1. (STMI<sup>R</sup> computation)**

1. **begin**
  2.     **computation of clean MTF**
  3.     **for** ripple pair  $\{\alpha, \beta\}$
  4.          $S^{100\%}(t, x) \leftarrow L(1 + 100\% \sin(2\pi(\omega_\alpha t + \Omega_\beta x) + \psi))$  (modulated ripple)
  5.          $S^{0\%}(t, x) \leftarrow L(1 + 0\% \sin(2\pi(\omega_\alpha t + \Omega_\beta x) + \psi))$  (flat ripple)
  6.          $y_{\alpha, \beta}^{100\%}(t, x) \leftarrow \text{auditory-spectrogram}(S^{100\%})$
  7.          $y_{\alpha, \beta}^{0\%}(t, x) \leftarrow \text{auditory-spectrogram}(S^{0\%})$
  8.         **for** cortical filter  $(i, j)$  (cortical modulation filterbank)
  9.              $r_{\alpha, \beta}^{i, j}(x; 100\%) \leftarrow \int_T \|y_{\alpha, \beta}(t, x; 100\%) *_{t, x} STRF^{i, j}(t, x)\| dt$
  10.             $r_{\alpha, \beta}^{i, j}(x; 0\%) \leftarrow \int_T \|y_{\alpha, \beta}(t, x; 0\%) *_{t, x} STRF^{i, j}(t, x)\| dt$
  11.             $R_{\alpha, \beta}^{i, j}(x) \leftarrow r_{\alpha, \beta}^{i, j}(x; 100\%) - r_{\alpha, \beta}^{i, j}(x; 0\%)$  (subtract baseline)
  12.             $MTF(x; \omega_\alpha, \Omega_\beta) \leftarrow \frac{1}{|\hat{w}| \cdot |\hat{w}'|} \sum_i \sum_j R_{\alpha, \beta}^{i, j}(x)$
  13.         **repeat algorithm for noisy MTF\***
  14.          $STMI^R \leftarrow 1 - \|MTF - MTF^*\|^2 / \|MTF\|^2$
  15. **end**
- 

The algorithm uses both fully modulated ( $S^{100\%}$ ) and flat ( $S^{0\%}$ ) ripples for MTF calculation. The flat stimulus was used as an estimate of the base level of the signal, and thus its content is subtracted from the 100% signal, leaving only the modulation content present above the baseline level. The modulation energy at each cortical filter characterizes the model's expected response to the ripple signal. Hence, the STMI<sup>R</sup> was estimated as a global measure of the attenuation in the modulation transfer function of the channel. This eventually translates to a measure of the expected intelligibility of a speech signal

transmitted through this channel.

Since the  $STMI^R$  is analogous to the traditional  $STI$  (with narrow-band carriers), we can compare the estimates of the former to those of the  $STI$  under distortions of white gaussian noise and reverberation. These noise conditions are the same distortions for which the  $STI$  has been reported to be a good predictor of intelligibility [84].  $STMI^R$  values were computed from clean and degraded modulation transfer functions (MTF and MTF\*) using algorithm 1. The results are displayed in Figure 4.2(A) for the reverberant and noisy conditions. The stationary white noise condition used here were generated by adding to the original signal a random Gaussian signal whose amplitude is defined according to the signal-to-noise ratio (SNR) level. The reverberation effect were produced by convolving the signal with Gaussian white noise whose envelope is exponentially decaying with various time constants.

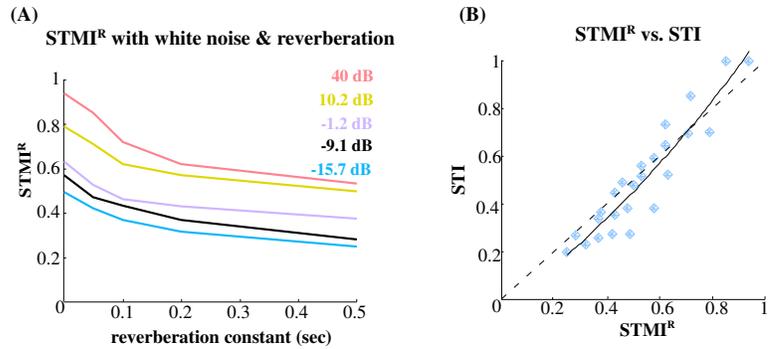


Figure 4.2: Effect of combined white noise and reverberation on  $STMI^R$  and  $STI$ . **(A)** The  $STMI$  values shown in this plot are computed according to algorithm 1 for noise conditions combining stationary noise and reverberation. **(B)** The panel shows the correspondence between the  $STMI^R$  and  $STI$  for the same conditions as in (A).

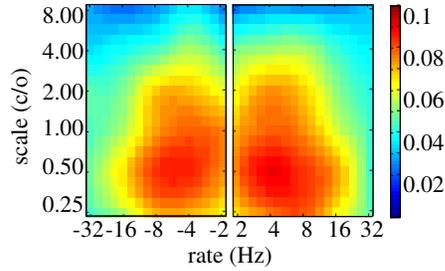
As expected, the  $STMI^R$  decreases with increasing noise and reverberation. We also measure the  $STI$  estimates under the same noise conditions, using the method derived by Steeneken and Houtgast [134]. Although different in details,  $STMI^R$  and  $STI$  measures

deteriorate similarly under these noise and reverberation conditions, and an approximate mapping between these two measures can be derived as shown in 4.2(B), exhibiting a very good correspondence between the two measures.

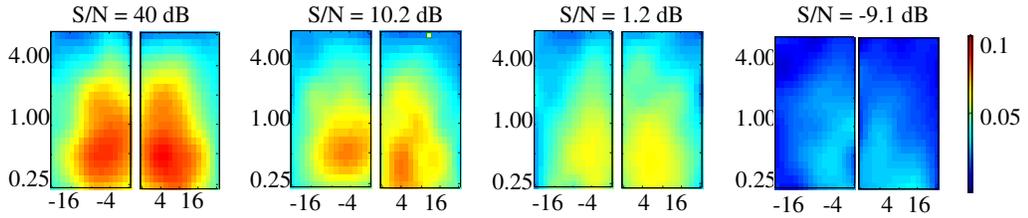
#### 4.2.2 Modulation Transfer Functions in noise

As the  $STMI^R$  captures the degradation in modulation content of ripples and its effect on intelligibility, it is valuable to examine the modulation transfer functions themselves. They are effectively quantitative descriptions of specific noise effects on particular modulation parameters. These MTFs are the basic computational parameter used to measure the  $STMI^R$  value, as detailed in algorithm 1. As different noise conditions are evaluated, the extent of the degradation is dependent on the rate and scale of the modulations, and the spectral content of the signal. Figure 4.3 summarizes the effect on all ripples (and the clean MTF of the auditory model in Figure 4.3(A)) of added white noise, different levels of reverberation, and combined effects of white noise and reverberation. In each case, the  $MTF^*$  is plotted as a function of  $\{\omega_i, \Omega_j\}$ , i.e., we integrate across the frequency axis  $x$ . It is important to note here that one can apply any arbitrary noise condition and compute the resulting  $MTF^*$  using exactly the same expressions presented in algorithm 1. These plots illustrate the effects of each of these distortions as follows. For white noise (Figure 4.3(B)) , the  $MTF^*$  is gradually and equally attenuated over all ripples. For increasing reverberation (Figure 4.3(C)) , higher rate ripples are more severely attenuated than lower rates. Both these trends are seen in Figure 4.3(D) for the combined noise and reverberation conditions. Note that the “random” weak patterns seen in Figure 4.3(D) reflect the random noise structure in a given trial, and hence are variable over different trials.

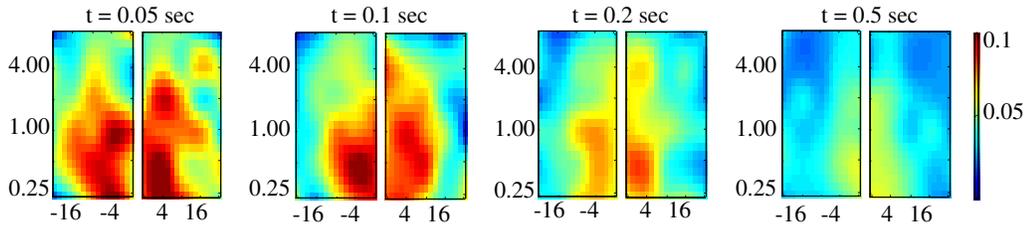
(A) MTF of auditory model



(B) White noise



(C) Reverberation



(D) White noise & reverberation

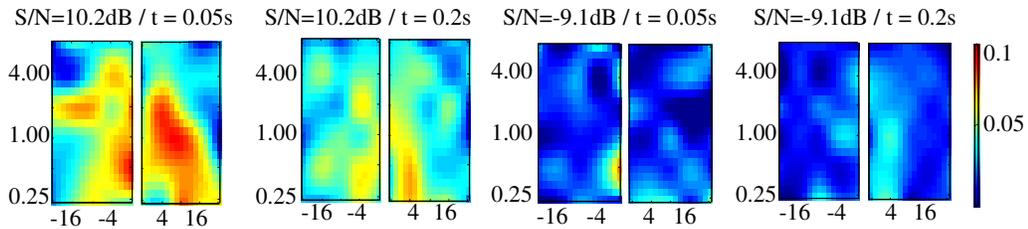


Figure 4.3: Effect of white noise and reverberation on the global MTF. (A) The global (clean) MTF of the auditory model computed from all ripples, summarized by the rate-scale plot (i.e., collapsing the frequency axis  $x$ ). (B) The attenuation of the global MTF (rate-scale plot) with increasing levels of white noise. (C) The attenuation of the global MTF at higher rates with increasing reverberation. (D) The combined effect on the global MTF of both additive white noise and reverberation.

### 4.3 Intelligibility for noisy speech

In many situations, the communication channel under study is often inaccessible, or we only have access to a noisy recording that was acquired in some unknown environment. The applicability of intelligibility metrics such as  $STMI^R$  is hence impossible for these cases motivating the need for an intelligibility metric directly applicable to noisy speech utterances or pre-recorded sentences. The following section describes an STMI variant called  $STMI^T$  (T, for templates) which addresses this case. We present the algorithm for  $STMI^T$  computation and validate its results by comparison to human scores reported by subjects in the context of psychoacoustic tests.

#### 4.3.1 $STMI^T$ procedure

In this section, we illustrate how the STMI is used to characterize the integrity of the spectrotemporal modulations of a given speech signal when distorted by various kinds of noise. We call this intelligibility metric  $STMI^T$ . Like the  $STMI^R$ , the  $STMI^T$  is also a measure of the integrity of spectrotemporal modulation content of a speech sample contaminated by noise. The  $STMI^T$  itself comes in two variants: it can be computed either with prior access to the original clean signal, or when only a noisy recording is available. In the first case, we can use the original clean sample as a reference (in a similar fashion as the clean MTF for  $STMI^R$ ) and follow the procedure shown in 4.1. In the latter case, we build a database of clean speech templates. This template database acts as general representation of the modulation content in any “clean” speech, irrespective of speaker, gender, linguistic material and content, and is hence a reference to which we can compare any noisy signal. We typically used about 200 sec worth of clean speech taken from utterances of the TIMIT speech database [1].

The  $STMI^R$  is in a way related to the speech-based standard STI, which uses the clean speech modulations as reference [121]. However, unlike the STI method, the  $STMI^R$  does not have to be limited to a specific clean reference, but uses generic templates of clean speech without requiring that the clean and degraded speech samples originate from the same exact tokens (same talker, same linguistic materials). Additionally, the STMI maintains its advantage over STI and classic intelligibility measures by robustly predicting sound degradation as result of severe and nonlinear distortions as we shall demonstrate.

The  $STMI^T$  computation is detailed in algorithm 2. The procedure consists of processing both the test signal  $x(t)$  and clean speech template(s) through the auditory model to assess their spectrotemporal modulation content. The  $STMI^T$  is then an estimate of the degradation of this modulation content between the clean reference and the test signal. Certain issues with the implementation of the algorithm can be addressed as follows:

- **Analysis frame:** The analysis of the noisy signal is performed over different frames of typically 2 seconds. Since the STMI procedure averages over the time dimension, we use a frame size that is short enough to sustain a valid stationarity assumption, but long enough to capture the temporal variability over a range of 1-2Hz. The same frame length is also used for the clean template(s).
- **Frame stack:** We use a stack of previously-analyzed frames of test signal (typically  $N=5$ ) to compare to the average clean template. The use of a stack of previous frames is aimed to give robustness to the STMI measure irrespective of the linguistic content at any particular frame. The choice of a relatively small stack size is however taken as a compromise between robust estimates and speed of intelligibility results.

- **Signal baseline:** Analogous to using a flat ripple in the STMI<sup>R</sup> procedure, we use a base signal to adjust the auditory outputs to the overall signal “spectrum”. The “base” is taken to be a stationary noise signal with a spectrum identical to that of the long-term average spectrum of the appropriate signal (clean or noisy speech).

---

♣ **Algorithm 2. (STMI<sup>T</sup> computation)**

1. **begin**
  2.     **initialize**
  3.          $L \leftarrow$  length of test signal  $x(t)$
  4.          $\tau \leftarrow$  frame size
  5.          $T \leftarrow$  Templates of “clean” speech patterns (over frames of length  $\tau$ )
  6.          $\Gamma \leftarrow$  Stack to hold  $N$  patterns of test signal
  7.          $i \leftarrow 0$
  8.         **do**  $i \leftarrow i + 1$
  9.              $F(t) \leftarrow$   $i^{\text{th}}$  non-overlapping frame :  $\forall t \in [\tau i, \tau(i + 1))$
  10.              $F(t) \leftarrow (F - \mu_F)/\sigma_F$  (normalize signal S)
  11.              $F_0(t) \leftarrow \text{fft}^{-1} [|S|e^{j2\pi\Theta}]$  (make base signal)
  12.              $y_F(t, x) \leftarrow$  auditory-spectrogram( $F(t)$ )
  13.              $y_{F_0}(t, x) \leftarrow$  auditory-spectrogram( $F_0(t)$ )
  14.              $r(\omega, \Omega; x) \leftarrow \int \|y_F(t, x) *_{t,x} STRF(x, t; \omega, \Omega)\| dt$
  15.              $r_0(\omega, \Omega; x) \leftarrow \int \|y_{F_0}(t, x) *_{t,x} STRF(x, t; \omega, \Omega)\| dt$
  16.              $r_i(\omega, \Omega; x) \leftarrow r(\omega, \Omega; x) - r_0(\omega, \Omega; x)$  (subtract base signal)
  17.              $\Gamma \leftarrow r_i(\omega, \Omega; x)$  (add to test signal stack)
  18.              $X \leftarrow \sum_{j=1}^N \Gamma_j$  (average signal stack)
  19.              $STMI^T \leftarrow 1 - \|T \cdot (X - T)\|^2 / \|T\|^2$
  20.         **until**  $i \leq \lceil \frac{L}{\tau} \rceil$
  21. **end**
-

As the STMI of a channel gradually decreases, speech transmitted through it should exhibit a concomitant loss of intelligibility that can be experimentally measured as increased phoneme recognition error rates. To relate the STMI values directly to experimental measurements of speech intelligibility, we plot in Figure 4.4(A) the  $STMI^T$  of speech tokens (computed from algorithm 2) with increasing additive noise and reverberation distortions.

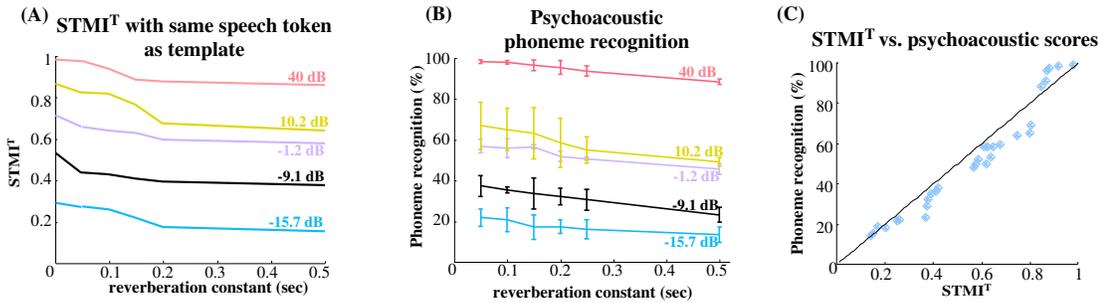


Figure 4.4: Comparing the effect of combined white noise and reverberation on the  $STMI^T$  and speech intelligibility. **(A)** The  $STMI^T$  of speech signals distorted by noise and reverberation. The  $STMI^T$  is computed according to algorithm 2 using same speech as template. **(B)** Experimental measurements of correct phoneme recognition of human subjects in noisy and reverberant conditions. **(C)** The  $STMI^T$  vs. correct percentages of human psychoacoustic experiments for the noise conditions given in (A) and (B).

### 4.3.2 Human psychoacoustic testing

Next, we performed a psychoacoustic test of intelligibility with four human subjects for white noise and reverberation conditions. Individual tests lasted about 3-4 hours, where the subject was presented with 240 sets of noise-contaminated speech samples. Each set consisted of five different CVC words (Consonant-Vowel-Consonant) played through a loudspeaker in an acoustic chamber. The subject was then asked to report all the phonemes heard through the loudspeaker. Afterwards, we counted the number of correct

phonemes reported by each subject, and averaged the scores of all subjects for each noise conditions. Figure 4.4(B) shows the error rates reported by the subjects in the psychoacoustic tests. Figure 4.4(C) illustrates the equivalence between the  $STMI^T$  estimates and the percent correct recognition scores found in these psychoacoustic experiments. The good correspondence between the  $STMI^T$  and the human scores confirms that the  $STMI^T$  is indeed a direct measure of intelligibility of noisy speech under conditions of combined white noise and reverberation.

As a side observation, one can notice the difference of the slopes of the  $STMI^R$  and  $STMI^T$  estimates in figures 4.2(A) and 4.4(A). Even though they both seem to have similar trends (drop with increasing reverberation level), the slope of  $STMI^T$  appears to be more shallow with reverberation, possibly because of time-averaging of the output patterns. Additionally, the two measures ( $STMI^R$  and  $STMI^T$ ) are conceptually different (use of ripples transfer function vs. speech tokens), and hence the difference in their trends. We can empirically account for the difference between these two measures by a simple expansive sigmoidal function.

## 4.4 Nonlinear speech distortions

What is the real advantage of the STMI over pre-existing intelligibility metrics? The STI has been widely and successfully used in speech intelligibility assessments under noise and reverberant degradation, and has also been adapted for use with speech signals directly [121]. Therefore, the results described above only demonstrate the correspondence between the STMI and STI, and hence the validity of the new STMI index. Here, we compare the performance of the two metrics under more difficult types of degradations: random phase-jitter and phase-shifts. We also include the results of psychoacoustic exper-

iments measuring the loss of intelligibility experienced by four subjects listening to words distorted by these two conditions. All psychoacoustic experiments were conducted exactly as described earlier in section 4.3.2. The subjects were presented with 160 different distorted words and were then asked to repeat the phonemes heard. Scores of average correct phonemes reported are presented in figures 4.5(B) and 4.6(B) for the two conditions.

### Phase jitter

First, we analyze a noise condition called *phase jitter*. It is a distortion commonly associated with telephone channels and caused by fluctuations of the power supply voltages [16, 95]. Communication engineers report that channels cannot be defended against such degradation, but it must be taken into account in the design of the receiver [95]. Therefore, studying the effect of this distortion on speech intelligibility is critical for improving the channel and receiver designs.

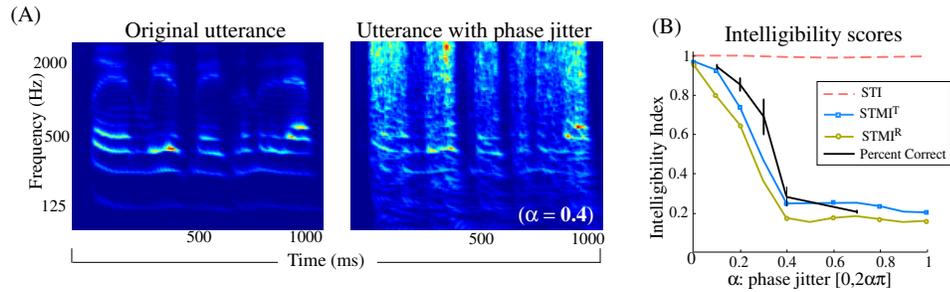


Figure 4.5: Effect of phase jitter on STMI and STI. **(A)** A clean utterance (left panel) is distorted through a phase jitter channel, with jitter variable  $\alpha=0.4$ . The distorted spectrogram in the right panel illustrates that the time dynamics of the signal are maintained while the spectral modulations are strongly affected by this type of noise. **(B)** The  $STMI^T$  and  $STMI^R$  drop as the jitter  $\alpha$  increases; while STI fails to capture the presence of noise in this channel. Scores of human listeners (black curve) show a good correspondence to the STMI trends.

Phase jitter is commonly modelled by the transform:

$$r(t) = \Re\{s(t)e^{j\Theta(t)}\} = s(t) \cos(\Theta(t)) \quad (4.1)$$

where  $s(t)$  is the transmitted signal,  $r(t)$  is the received signal, and  $\Theta(t)$  is the phase jitter function modelled as a random process uniformly distributed over  $[0, 2\alpha\pi]$  ( $0 < \alpha < 1$ ). The jitter effectively destroys the carrier of the speech signal leaving its envelope largely intact, especially for large values of  $\alpha$ . Though the temporal dynamics of the signal are not distorted, its spectral modulations are greatly affected by the phase jitter (Figure 4.5(A)). For  $\alpha = 1$ , the speech signal becomes a modulated white noise with the same envelope as the original signal. Figure 4.5(B) illustrates the expected loss of intelligibility as a function of jitter severity ( $\alpha$ ) as measured by the  $STI$ ,  $STMI^R$ , and  $STMI^T$  (computed as the mean of 10 different speech sentences from the TIMIT database). Both  $STMI$  measures deteriorate with increasing jitter  $\alpha$ . In contrast, the  $STI$  is *insensitive* to such distortion, and actually predicts an intelligibility of 100% even at extreme jitter conditions ( $\alpha = 1$ ), as shown by the dashed red curve. The failure of the  $STI$  can be directly explained by the fact that phase jitter primarily affects the spectral dimension in speech, and hence does not affect the modulation amplitude of the narrow-band carriers used in the  $STI$  measurement. The  $STMI$ , on the other hand, captures successfully the substantial effect of jitter on the spectrogram of oriented ripples as well as actual speech, and hence is able to correctly predict intelligibility loss. The results of human testing confirm this finding, as shown by the close agreement between human scores and  $STMI$  intelligibility estimates in Figure 4.5(B).

## Phase shift

Next, we analyze the noise condition of *inter-channel phase-shift* or *delay scatter*. This type of channel distortion is a linear phase-shifting of signal frequencies over limited spectrum ranges. Specifically, this distortion de-synchronizes frequency channels in the range 400-1900 (Hz) with respect to each other; by applying a simple linear phase shift that varies from one frequency band to the other. The actual phase shifts are defined by the function  $\Phi = \omega\tau_i$ , applied over 300 Hz frequency bands, each indexed by  $i$ , over the range 400-1900 Hz ( $i = 1, 2, \dots, 5$ ), where  $\omega = 2\pi f$  is the frequency at which the phase-shift is applied, and  $\tau_i$  is a parameter which controls the slope of the phase function in the  $i^{\text{th}}$  band. The effect of the phase-shifting appears as a spectrally jittered signal, but with minimal change to the temporal envelope modulation patterns at any given channel (Figure 4.6(A)). The parameter  $T$  captures the average time delay between the different frequency bands, and controls the severity of de-synchronization.

Figure 4.6(B) illustrates the decrease in  $STMI^R$  and  $STMI^T$  with increasing delay scatter (over a range of values  $T$ ), consistent with the increasing channel distortion of the spectrogram of the ripple and speech signals. The  $STMI^T$  drops faster because of the specific arbitrary choice of frequency bands and shifts; and the drop (while it always occurs) is variable in steepness depending on the exact test utterance. As with the previous phase-jitter distortion, STI measures (noise or speech-based) are expected to be insensitive to such phase-shift because this distortion does not significantly affect the modulated envelope of the narrow-band carrier test signals used in standard STI computations, nor does it affect the envelope modulations of the speech spectrogram. Human intelligibility tests exhibit the same deterioration as that predicted by the STMI as illustrated in Figure 4.6(B). Our results are comparable to those of Greenberg and Arai [74] who studied the

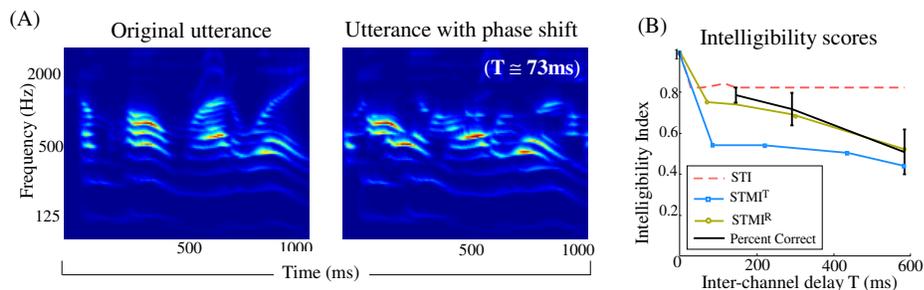


Figure 4.6: Effect of linear phase shift on STMI and STI. **(A)** The input speech signal (left panel) is distorted by linear phase shifts on different frequency bands. Five frequency bands (of uniform 300 Hz ranges going from 400 to 1900 Hz) are phase-shifted according to the vector  $[\tau, 2\tau, -3\tau, 4\tau, 5\tau]$  (a different phase shift per frequency band) where  $\tau$  is the parameter that controls the amount of shift per band. The result is a de-synchronization of the different frequency bands relative to each other, as shown by the noisy spectrogram in the right panel. The shift parameter used in this case is  $\tau=0.5$  (average time shift across channels of  $T \cong 73\text{ms}$ ). **(B)** The effect of the shift parameter  $\tau$  (or  $T$ ) on the  $\text{STMI}^R$ ,  $\text{STMI}^T$  and STI. Since for each value of the shift parameter ( $\tau$ ) different frequency bands are time-shifted with various amounts relative to each other, the x-axis of the graph gives an average estimate ( $T$ ) of the time shifts across the different frequency bands (where  $T \cong 146\tau(\text{ms})$ ). The black curve shows the intelligibility scores of human listeners tested with the same phase shift conditions.

intelligibility of a similarly (but not identically) distorted speech and concluded that scores dropped below 50% only after the channel jitter exceeds 200ms.

## 4.5 Conclusion

In this chapter, we presented a quantitative assessment of intelligibility based on the modulation content of signals. The algorithms developed here were based on a biological foundation relating modulation content to speech intelligibility. While quality of sound ultimately relies on the exact temporal patterns of a signal, understanding speech, despite its lack of natural quality relies solely on its modulation components.

The key findings presented here can be summarized as follows: **(1)** We have shown a direct mapping between speech intelligibility and spectrotemporal modulations of speech, **(2)** we have defined a new intelligibility metric (STMI) that has proven its robustness under severe and nonlinear distortions, hence superseding previously available intelligibility metrics (e.g. the speech transmission index, STI), **(3)** we have defined a biological framework explaining the superiority of the STMI, as well as explaining the benefits and limitations of the STI, and finally, **(4)** we have refined our understanding of the role of receptive field patterns in the cortex (particularly STRF<sup>E</sup>s). Earlier in chapter 3, we have defined the functional significance of the presence of representations of modulation envelopes in cortical responses, to the fact that slow modulations are the main carrier of information in speech. The predominance of such STRF<sup>E</sup> patterns seems in accord with the role of cortical circuitry in contributing to sound perception and understanding.

## Chapter 5

# Auditory scene analysis

As the brain takes on the job of representing the world that surrounds us, the auditory system is responsible for making sense of the world of sound, in a process referred to as *auditory scene analysis*. Technically, the term auditory scene analysis refers to the cognitive task engaging the auditory system in identifying and building auditory objects from mixtures of sounds present in a complex acoustic environment. In other words, the auditory system is responsible for identifying how many sounds are present in the environment, where they are coming from, and what do they mean. The spectrogram in Figure 5.1 shows a mixture of two sounds: a male speaker uttering the phrase */she had your dark suit/*, when a cello starts playing a note about 750ms into the sentence. While we know now that the spectrogram in the figure corresponds to these two particular sounds, the auditory system has no prior knowledge of what the incoming sound mixture

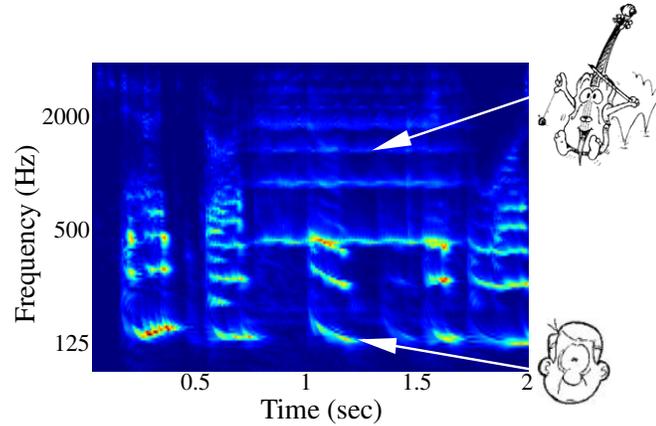


Figure 5.1: Spectrogram of a sound mixture. A male speaker is uttering the phrase */she had you dark suit/* while a cello note is being played.

consists of, how many sounds are there and when each sound will begin and end. A similar challenge is also faced by computational systems trying to mimic the auditory functionality in sound separation and auditory scene analysis.

When talking about auditory scene analysis, a common term that gets quoted quite often is the word “stream”. A formal definition of this term was given by Bregman ([21], pg. 10): *“Our mental representations of acoustic events can be multi-fold in a way that a mere word ‘sound’ does not suggest... It is useful to reserve the word ‘stream’ for a perceptual representation, and the phrase ‘acoustic event’ or the word ‘sound’ for a physical cause”*. Hence, we use the term stream to describe an internal representation of what would be perceived as a distinct auditory object. Our computational strategy aims to represent an acoustic mixture into various distinct auditory objects following principles dictated by perceptual grouping and auditory organization.

Typically, there is a distinction in the literature between two perceptual components of auditory scene analysis: *auditory streaming*, which refers to the perceptual organization of sounds in time, and *sound separation*, which deals with concurrent sound segregation. While not strictly independent, these two concepts are generally studied separately. On

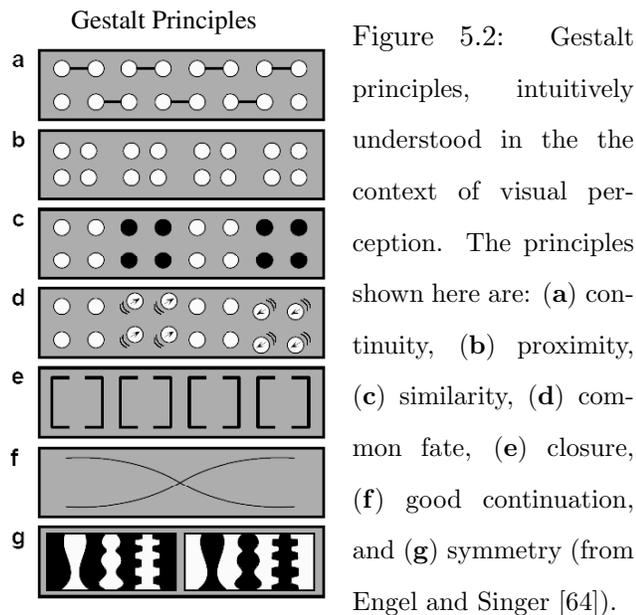
the one hand, interest in auditory streaming comes from psychologists and physiologists studying the neural and cognitive basis of auditory organization. Musicians are also widely interested in this idea, as melodies tend to exploit the brain’s ability to stream sounds in time. Baroque melodies for instance tend to introduce *multiple* streams emanating from one *single* instrument, by rapidly switching melodic sequences. On the other hand, studies of sound separation rise *mostly* from the artificial intelligence and engineering community. People interested in computational models of scene analysis are primarily interested in building intelligent systems that can “hear” (i.e. identify and segregate sound elements). Such systems, for the most part, are applied as front-end improvements for automatic speech recognition and speaker identification models.

In this work, we present a modelling scheme of auditory scene analysis based on principles of cortical sound processing, where we address various components pertinent to both auditory streaming and sound separation. The chapter starts by a review of the general principles of auditory perception, as well as an account of the available models in the literature addressing the scene analysis problem. The following section lays the basis for our proposed approach by motivating the choice of both unsupervised learning, and Kalman-based estimation to achieve an adaptive prediction-based system. We then elaborate on the implementation of this model in the context of a multi-scale multi-rate cortical scheme. Next, we present a series of simulations addressing both the auditory streaming problem, as well as speech separation cases; and conclude by a assessment of some insights into biological auditory scene analysis learned from this model.

## 5.1 Perceptual principles

### 5.1.1 Gestalt principles

The foundation of most work in auditory scene analysis can be traced back to early work in perceptual psychology from the late 19th century. Originating from work by Ehrenfels (1890) and Mach (1886), these principles (referred to as *Gestalt theory*) explain our psychological, physiological, and behavioral perceptions in their contextual framework, in accordance with a number of simple principles [21]. Specifically, they address the binding question in our perception (auditory, visual ...), and how sensory input elements are combined together to create mental patterns, or sensations of individual objects in the environment. The German word *Gestalt* means “pattern” or “shape”, hence the use of this term in reference to how sensory components are grouped together to create mental patterns.



Gestalt theories have mostly been explored in visual perception, where they made an important and lasting contribution to our understanding of the vision problem. Most

Gestalt principles are in fact easily described in their visual context, as shown in Figure 5.2. Similar principles can also be described in the context of auditory events. Their application for auditory perception is interpretable in terms of grouping cues used for combining together auditory elements that belong to a common sound stream or percept. We elaborate on five of these principles, as they are the most frequently cited as playing a perceptual role in parsing the auditory scene into individual streams:

1. *Proximity*: Elements that are close together in time or space tend to group together.

In Figure 5.2(b), the quadruplet sets of individual circles tend to be perceived as independent sets (the figure contains four sets), since the circles within each block are closer together than circles from different blocks. In the context of auditory events, this principle is used to refer to distances between physical acoustic attributes such as frequency, onset, pitch and loudness. Frequency proximity is a straightforward illustration of this principle. Frequency separation between sounds affects their perceptual coherence, leading sounds that are close in frequency to group together [21].

2. *Similarity*: Elements that lie closely together group together. The similarity in this sense refers more to characteristics of the object or element. In case of visual objects, it includes characteristics such as shape, size, color, texture, value, etc. In Figure 5.2(c), though all circles are equidistant, those similar in color tend to be perceived together as belonging to one block. In the context of auditory perception, this principle refers to acoustic features that cannot be described with a single physical attribute, such as similarity in timbre [76]. The principles of “similarity” and “proximity” are hard to distinguish from each other in audition, and do basically lead to very similar effects. Bregman suggests to reserve the use of the term “similar-

ity” to cases where it is not possible to describe the physical basis of the dimension along which two auditory elements are similar, such as similarity in timbre, instead of proximity in timbre [21].

3. *Good Continuation*: Elements that maintain a smooth variation in time or space tend to group together (Figure 5.2(f)). In terms of sound attributes, it refers to the continuity or flow-in-time of sound originating from a common source. Intuitively, abrupt changes indicate the appearance of a new source. Grouping of sounds is in fact enhanced when frequency changes preceding and following interruption by noise or silence are consistent with each other [21]. In this sense, “good continuation” can be thought of as the continuous limit of the “proximity” principle.
4. *Closure*: Elements that form enclosed objects group together (Figure 5.2(e)). This principle is very much invoked in visual scenes in cases of occlusions, which are similar to the problem of masking in auditory scenes. The principle of “closure” is in fact interpreted as an auditory compensation for masking [76], and evokes the completion of fragmented sound features, as in the case of the continuity illusion. This phenomenon occurs when a sound is briefly interrupted by a gap of silence. If the gap is filled by a noise burst, the original sound is perceived as continuous through the noise distracter. The *illusion* that the sound continues through the noise occurs only when the frequency content and timing of the noise coincide such that it is plausible that the sound would have continued smoothly through the noise burst. This principle is commonly evoked for speech perception in noisy environment. Carlyon and colleagues showed two-formant vowels can be easily identified when alternating the formants in time while filling that gap silences with noise [28].

5. *Common fate*: Elements with synchronized similar variability tend to group together (Figure 5.2(d)). In the auditory context, this principle states that sound elements with synchronous changes such as common onset, vibrato, etc, fuse together as emanating from a single source. The main two type of auditory synchronous changes studied in the literature are AM (Amplitude-Modulation) and FM (Frequency-Modulation) changes [21].

We shall review the acoustic correlates of these principles in the following section. By translating Gestalt principles into concrete grouping cues, it is possible to give a precise formulation of their perceptual manifestation and their computational role, and hence lay the foundation for formulating a model of auditory scene analysis inspired from these perceptual principles, and guided by their underlying neural mechanisms.

### 5.1.2 Acoustic correlates

Psychoacoustic studies of auditory streaming have been accumulating evidence on grouping cues for many decades now. An exhaustive list of auditory cues known to play an important role in auditory grouping and streaming of sounds has been compiled by Bregman in his “Auditory Scene Analysis” book [21]. Grouping cues are mostly acoustic correlates of one or more Gestalt principle described above. They determine whether sound components are to be fused together into a single perceptual stream or segregate into separate streams. While it is becoming more evident that any sufficiently salient perceptual difference along *any* auditory dimension may lead to stream segregation [110], we review here various factors that are frequently mentioned in the literature for their role in sound segregation or fusion.

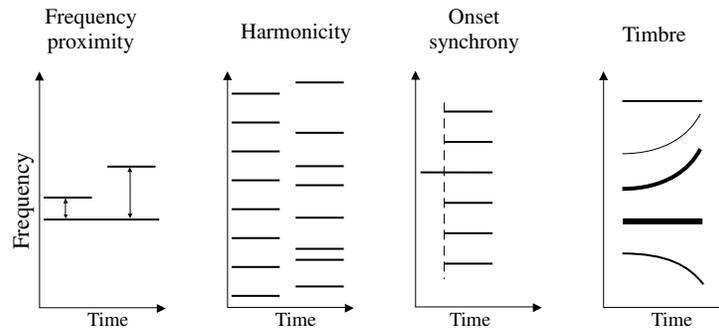


Figure 5.3: Auditory grouping cues. *Frequency proximity:* Tones that come close together in frequency group together. *Harmonicity:* Frequency channels that are harmonically related fuse together. *Onset synchrony:* Frequency channels that share a common onset group together. *Timbre:* The frequency channels that move up together maintain their spectral spacing and hence group together. The other frequency components exhibit a different timbre quality and hence separate as different streams.

### Frequency separation

Frequency separation is one of the most commonly invoked cues when talking about auditory grouping. Sounds that are close together in frequency fuse together (Figure 5.3). But how close do sound components need to be? A leading theory among experimental psychologists and psychoacousticians addressing this particular point is *peripheral channelling* [110]: Streams are believed to fuse together when their excitation patterns in the auditory periphery overlap. In other terms, grouping effects are promoted if the frequency contents of two streams fall into the same cochlear channel. A classic study reviewing the physical parameters of auditory streaming was performed by van Noorden in the 1970s [140]. He explored the role of frequency proximity and presentation rate in streaming an alternating sequence of tones “A” and “B” alternating in a sequence ABAB... As the tone frequencies were brought closer together below a certain interval, listeners reported hearing one stream instead of two. Van Noorden’s experiments set the ground for most of the subsequent work in auditory streaming using alternating sounds. Studies have extended

his findings to other factors which also influence streaming, such as temporal envelopes and spatial location of sounds [37, 110]. Non-alternating sounds were also used for testing the role of spectral separation in stream segregation. Hartman and Johnson [77] tested the idea of peripheral channelling by using an interleaved melody paradigm. Psychoacoustic tests were performed for melody identification as streams of different melodies were manipulated into different cochlear channels. As expected, a successful segregation of the melodies was achieved when the streams were shifted into separate peripheral channels.

### **Harmonicity**

Frequency channels that stand in harmonic relationship to each group together. It is in fact known that harmonically related (or closely harmonic) complexes fuse together into a single pitch (Figure 5.3). Based on this principle, complex sounds can be distinguished based on the relationship between their fundamental frequencies. In a study by Rasch in 1978 [124], he demonstrated that as the fundamentals of two complex tones depart from simple harmonic relationship, the complex tones are heard more clearly as distinct entities. Sound sources are in fact discernible as soon as their F0 difference is detectable, and particularly in the case when their harmonics are resolved [42]. Frequency channels do not have to be *only* harmonically related before they can fuse together. Work by Roberts and Bregman [126] as well as others has provided evidence of fusion due to regularity in spectral spacing. This shall be mentioned again in the section on the role of timbre as a grouping cue.

### **Onset/Offset synchrony**

Onset and offset synchrony are also major grouping cues. Sound components that start or end together in time are likely to have originated from the same sound source (Figure 5.3). Rasch showed that two concurrent tone complexes are perceptually more distinct when the sounds start at different points in time. An onset asynchrony as small as 10-30ms increases the perceptual saliency of the tones in a mixture [21, 124], but our ability to detect the difference in time of arrival of energy across different channels is in fact a little as 1-2ms [42]. Offset asynchronies appear to also play a similar role as onsets, although their effects were reportedly less pronounced [42].

### **AM and FM modulations**

Temporal regularities in sound components - in the sense of common fluctuations of sound both in frequency and amplitude, are important grouping cues ([72] and references therein). As sounds are amplitude modulated, the auditory system uses differences in their amplitude trajectories to distinguish the number of sources present and segregate the different streams. Early work of von Békésy in the 1960s showed that modulating tones at rates below 50 Hz helped listeners fuse together those that share a common modulation. Various studies have later confirmed and extended such finding, as most available evidence focuses on the role of slower fluctuations [42]. Some studies have attempted to explore the role of faster variations rates. An interesting study by Grimault and colleagues [75] has in fact demonstrated a clear role of envelope fluctuations in stream segregation -in the absence of any spectral or temporal fine structure cues, and seemingly independent of modulation strength. On the other hand, the role of FM modulations in stream segregation is still ill defined, with highly inconclusive results. FM incoherence is particularly

hard to distinguish from harmonic mistuning. Nonetheless, there is some evidence that FM contributes to the prominence of fusion of frequency components particularly in case of masked vowels, or possibly in the presence of a steady noise background [42].

### **Sound timbre**

Timbre is less frequently mentioned as a possible cue for auditory streaming. It is a multidimensional percept influenced by spectral shape, envelope shape, and spectral changes over time [110]. It is now established that the auditory system exploits the regularity of the spectral ratios of a pattern to fuse sound components together (Figure 5.3). A vibrato produces a frequency modulated sound complex, whose partials move up and down in synchrony with each other, and hence fuse together as a coherent stream. Studies investigating the timbre cue were mainly interested in addressing the role of peripheral channelling in stream segregation, which tend to relate the timbre cue with the frequency proximity. To explore the relevance of timbre regularity, a study by Dannenbring and Bregman [40] addressed factors inexplicable by the channelling theory. Dannenbring and Bregman tested streaming effects using a tone-noise pair. The noise was constructed to be narrowband (about 1.5 semitones bandwidth), and maintaining its bandwidth to less than one ERB guarantees that it induces similar excitation patterns as a tone, and hence limiting any effect of peripheral channelling. The study found that segregation was greater for tone-noise combinations than tone-tone or noise-noise sequences. The separation was explained by the difference in temporal envelopes of tones and noise, hence leading to a separation in timbre. A summary of other studies exploring the role of timbre as grouping cue is reviewed by Moore and Gockel in [110].

## **Spatial location**

Sounds in natural environments emanate from different spatial locations. As they come into the ear, localization cues based on sound binaural irregularities can identify the location of sound sources. It is reasonable to think that sound elements originating from a single source would share common spatial cues, and hence would tend to perceptually group together. Surprisingly, evidence from experimental psychoacoustics hints to a relatively weak role of lateralization in auditory grouping [21, 42]. It should be noted that quantifying the role of true lateralization and binaural mechanisms while neutralizing other acoustic grouping cues is quite challenging. Nonetheless, spatial location effects have been studied under particular testing conditions. Hukin and Darwin have suggested a strong effect of lateralization in vowel identification when the direction of sound is previously cued (i.e., attending to sound from a particular direction prior to the test signal) [42]. Yost *et al.* [149] has also demonstrated a more pronounced role of binaural processing in “cocktail party” situations when dealing with more than two concurrent sound sources.

## **Integration across cues**

A question that naturally arises from listing the various grouping cues is how do they compare to each other in terms of importance, and how does the auditory system incorporate the information provided by each one of them to segregate the different streams in an auditory scene. Psychoacoustical evidence indicates that grouping is not “all-or-nothing” [42]. The auditory system appears to gather evidence from all available sound features. It then chooses the most appropriate ones for segregating streams.

### 5.1.3 Top-down effects

Stream formation is a highly context-dependent process, giving rise to a very ambiguous definition of what a perceptual auditory stream is. In Bregman’s own words, the use of the word “stream” gives us the flexibility to “*load it up with whatever theoretical properties seem appropriate*” [21]. A stream is a “*perceptual unit that represents a single happening,*” where a happening can be the sound from a particular source, or even a collection of different sounds distinct from others in the environment. For example, there is no clear answer as to whether to call a melody played by an entire orchestra “a stream”, or whether the particular piano or violin tunes by themselves are what should be defined as “streams”. The particularity of the human brain is its ability to direct attention to the various sound sources in the environment, and to select the level of granularity at which to attend to sounds. The orchestral melody can be considered as a whole; or we can choose to focus our attention on the specific instrumental tunes, or even to the sounds from the audience. Naturally, our ability to attend to specific or multiple streams is limited by our brain’s cognitive capabilities. Actively listening to more than one conversation is very difficult if not in fact impossible.

The literature reviewing the role of attention in auditory streaming is very diverse and controversial, especially in formulating theories relating auditory stream formation to attention. There is no doubt that attentive mechanisms are invoked in auditory streaming [27, 41]. However, the specific relationship between primitive segregation and auditory selective attention is not yet resolved. Bregman favors an interpretation of stream segregation mediated by pre-attentive grouping mechanisms [21]. In his book, Bregman postulates that auditory scene analysis entails two complementary processes: (a) *primitive segregation*, which is a bottom-up pre-attentive process that parses a sound mixture on the

basis of its acoustic properties, and (b) *schema-based segregation*, which is a top-down process that involves learning and memory, and reflects our prior experience and familiarity with the linguistic material or sound pattern. Primitive mechanisms rely on the general properties of the sounds, and extract acoustic cues (as described in the previous section) which we innately use for parsing streams, without relying on any previous experience or knowledge. Schema-based mechanisms, on the other hand, invoke our specific experiences and familiarity with sound sources and materials. For instance, contextual expectations help us fill-in parts of a speech signal if masked or absent. They also help us recognize our names in a sound mixture much easily than other less familiar sounds.

## 5.2 Literature review of CASA techniques

Efforts to understand the computational basis of streaming and sound segregation led to a wide interest in building systems that perform “intelligent processing of sound mixtures” [37]. The field of experimental auditory scene analysis slowly gave birth to a growing body of work in computational sound analysis or what is commonly referred to as *Computational Auditory Scene Analysis* (CASA). Research in CASA has a strong interdisciplinary component, as it draws its basis from psychophysical and psychoacoustical theories of hearing, as well as machine learning and artificial intelligence principles.

Computational models for speech separation started with early work by Parsons in 1976 [120], who introduced one of the original models of speech separation based on pitch tracking in time. Attempts to build voice separation systems were followed by work of Weintraub in the 80s [146], who used the difference in fundamental frequency to separate two simultaneous voices. His system was based on coincidence detection (à la Licklider [97]) as a measure of periodicity in different frequency bands; augmented by a

Markov model for tracking the states of different speakers as they progress between voiced, unvoiced, silent and transitional states. The outcome of the model allowed a tracking a recovery of individual spectra of two simultaneous voices.

Attempts to give a more biological foundation to computational auditory scene analysis needed to address the various dimensions along which the biological system operates. Efforts in this direction were launched with work of Brown and Cooke [23, 24, 38] who introduced the notion of *auditory maps*. Their systems perform segregation in a purely data-driven manner. Incoming sounds are processed using an auditory “font-end”, whose output is segmented into atomic units. The acoustic elements are grouped into feature maps that organize the auditory cues along dimensions of harmonicity, common onset, continuity, modulation, etc. These are subsequently used to construct symbolic descriptions of the auditory scene. The pitch values, onset and offset times, etc, are used as grouping cues to cohere a group of auditory elements into their corresponding streams.

While the approach undertaken by Brown and Cooke tries to employ some biological realism into their computational models, schema-related information is intrinsically lacking in their strategy; an argument that was at the core of Ellis’s prediction-driven CASA system [63]. His approach sought to reconcile the observed acoustic features with the predictions of an internal model. At any time instant, adjustment of the model’s predictions are made based on its previous expectations from prior instants combined with evidence from external observations. The predictions are based on a set of pre-defined abstract elements: noise clouds, transient clicks, and wefts (representing wide-band periodic segments), and the input signal is mapped into elements based on this sound vocabulary.

Recently, neural network models have also been sought to construct models of auditory function. The theory of oscillatory networks was a particularly attractive one for

scene analysis, as different populations of neural oscillators can be used to represent perceptual streams. Work by Brown and Cooke [25], and Wang and Brown [143] presented a model of stream segregation based on oscillatory networks, where synchronized oscillators are interpreted as belonging to a common stream. Synchrony of individual oscillators is initiated by regularity in the sound's spectrotemporal elements, and hence lateral connections between oscillators are implemented to encode harmonicity, proximity in time and frequency.

On the other side of the spectrum, many studies have attempted to solve the problem of sound separation from a non-biological perspective. Systems built in this spirit are free to expand on any cues with no constraints limiting their use on grounds of biological plausibility. They are, however, limited by their own mathematical construction or model-based approaches. Such is the case for the widely used Blind Source Separation (BBS) techniques which rely on the statistical independence of the sources in the mixture [15]. As its name indicates, blind source separation attempts to separate a mixture of signals into their individual sources, with no prior knowledge about the source statistics in general (hence the term blind). BBS techniques are quite successful, but mostly when information from *multiple* sensors is available and statistical independence or prior statistical regularity is satisfied [150]. However, stream segregation continues to be a challenge for strictly statistical techniques like BBS systems, particularly in the monaural case.

If any criticism is to be addressed to these currently available systems, it would most likely be that their performance on the general task is rather poor. Most systems are, however, to be credited for exploring important aspects of sound analysis, and examining the role of many perceptual principles for the organization of sound. While their results apply reasonably well to specific tasks set by the investigators, the applicability of their

approaches to solving a general sound separation problem is quite limited, and the gap between CASA systems and real audition is still considerably wide. These systems fall short in applying successful information integration strategies to consolidate information extracted from the acoustic signal, with contextual facts from the environment. A major concern in most systems in fact deals with which patterns are to be extracted, and how successful their cue-detection schemes operate. While these models are very useful for a further understanding of sound organization principles, they lack a generalized scheme for acoustic pattern integration. A system such as that presented by Weintraub [146] for pitch extraction and tracking is not sufficient nor practical for building a robust sound separation system. Other systems trying to formulate a more general approach such as that undertaken by Ellis are constrained by a pre-defined classification of sounds into specific elements (noise, transients, etc) [37, 63]. Such limitation does not generally lead to an accurate characterization of acoustic events, hence resulting in very ambiguous outcomes.

Along with “general-purpose” models, more specific approaches have also been proposed which do not specifically tackle auditory scene analysis as a comprehensive scheme to building intelligent systems that can hear. Mostly relevant to our current study are models addressing the question of auditory streaming or sequential stream segregation. The earliest computational account for auditory streaming was presented in a model by Beauvois and Meddis [12]. This model attempts to give an explanation of streaming in terms of peripheral channeling. Streaming effects are simulated based on frequency differences and presentation rates of alternating pure tones, where frequency channels compete with each other to result in one dominant *foreground* stream. An elaboration of this model has been presented by McCabe and Denham [103], who extended it from a two-channel to a multi-channel model. This new approach also introduced inhibitory feedback loops that

enabled the model to allow for interactions between current and past response patterns in different frequency channels. The model was successfully applied to simulate various streaming effects such as alternating tone sequences, as well as the interaction between background and foreground perceptions, and the temporal buildup of streaming perception. Nonetheless, both models are limited by their dependence on peripheral channelling to account for streaming effects, and cannot easily generalize to general sound segregation situations, such as concurrent speech. The models are also likely to fail in simulating streaming effects that rely on segregation based on more complex grouping cues, such as spectral patterns or timbre features. A case of streaming ripple sequences (as will be shown in the results section of this chapter) is a particularly detrimental test that relies on streaming broadband noise sequences that only vary in their spectral regularity. Any model that relies solely on peripheral channelling or frequency separation is bound to fail in accounting for such effect as both sounds (the two ripples to be streamed) are in effect broadband, noise-like stimuli, that excite the same frequency regions along the tonotopic axis.

### **5.3 Adaptive ASA architecture**

Going back to the perceptual principles of sound organization, we have described a set of acoustic cues that give the auditory system evidence from the environment as to which elements should group together, and which should separate. The challenge that follows is how to actually organize these features together so that they can lead to perceptually meaningful streams. If we adhere to a modelling approach that does not rely on any dictionaries of linguistic knowledge or databases of familiar sounds, our only hope for achieving this goal is to rely on a robust representation of acoustic features that would

expose regularities in sound patterns. This regularity reflects the constraints imposed on acoustic events by the environment, as well as internal expectations arising from our internal model of the world. In this work, we propose a model based on interaction between this “higher-level” internal representation of the world and unsupervised sensory feature classification, all governed by principles of cortical sound processing.

### 5.3.1 Unsupervised learning

Based on a set of feature maps representing various auditory cues described above, the brain has to decide which sound elements cohere together into independent streams. In the absence of any external supervision or prior information about the sound mixture itself, the auditory system has to self-organize the sound elements into streams in an *unsupervised* (or self-supervised) fashion. Unsupervised techniques are neural network learning approaches that are notably biologically plausible. The biological system does not in general have prior knowledge on what kind of input will be impinging on its sensory organs, except for possible cases of expectation or context. And even in those cases, the specifics of the external input are still unknown to the system. In building a computational model mimicking the biological stream segregation process, the challenge lies in finding *natural groupings* for clustering the different patterns in the input.

The literature of unsupervised learning is quite rich, and involves a wide spectrum of techniques ranging from feature extraction methods that extract statistical regularities directly from the input, to density estimation that builds parameterized statistical models of the input. Somewhere in the middle of this spectrum lies competitive learning as an approach whose goal is to distribute a number of vectors in a possibly high-dimensional space. It can be implemented based solely on density estimation or feature extraction, but

is most often formulated as a combination of the two methods. Competitive learning is considered by many to be “a strong candidate model of cortical learning” [13], which makes it very appealing for the task at hand. Techniques used for competitive learning involve classifiers which for the most part operate on the basis of spatial or pattern regularity, to find natural partitions of the input’s high-dimensional space. While quite successful, most available techniques rely only on the low order statistics of their inputs. They handle the inputs as a set of independent identically distributed (iid) samples, then implement some form of Hebbian learning or probabilistic maximum likelihood estimation [17]. In doing so, they fail to take into account any smoothness or coherence constraints in the signal except in cases of hand-crafted architectures.

Hearing, however, is a dynamic process which tracks the changes in sound patterns as they evolve in time. The temporal continuity in sound streams is a major clue as to whether they come from a common source or not. Sound production (be it speech, music, nature, etc ...) obeys natural laws that dictate a certain degree of smoothness in the evolution of acoustic events in time. If a spectral pattern exhibits a sudden or unexpected discontinuity, it is highly likely to have been generated by a different sound source. Following these considerations, it is more appropriate to adopt an adaptive form of competitive learning that takes into account these smoothness constraints, and can build expectations of future responses based on inferences from the past and present.

### **5.3.2 Adaptive competitive learning**

Sound mixtures entering the ear reflect a temporal regularity imposed by environmental constraints and laws of nature. The role of the auditory system then is to make sense of these sounds, which in effect translates to estimating and predicting an internal (hidden)

state of the observed dynamical system (i.e, estimating the events in the environment that produce the sounds). Accurate recognition of the sound streams present in the mixture is synonymous to a correct estimation of the parameters of the model/environment giving rise to the mixture. Such approach conforms with the principles of statistical inference [85]. In effect, hidden (or “latent”) variable models are techniques that attempt to explain observed data by some underlying hidden factors that we only have indirect information about. Such models are well suited for incorporating a representation of the environmental constraints on sound streams, and for adaptive prediction of the model’s probabilistic representation of the observed environment. Such predictive approaches operate according to the following scheme: sensory information provides evidence to the system, which would either justify or cause it to change its internal representation of the world and the state of objects in it.

The model we consider here consists of a set of random processes  $\mathcal{Z}(t)$  symbolizing the auditory objects (or streams) present in the scene, which are unknown a priori. The set of  $K$  possible streams in the environment is represented by the vector set:  $\vec{\mathcal{Z}} = \{\mathcal{Z}^1, \mathcal{Z}^2, \dots, \mathcal{Z}^K\}$ , where each vector  $\mathcal{Z}^\alpha(t)$  is an internal (abstract) representation of stream  $\alpha$  at time  $t$ . The sound mixture (or observed inputs) is represented by a set of feature vectors:  $\vec{\mathbf{I}} = \{\mathbf{I}^1, \mathbf{I}^2, \dots, \mathbf{I}^L\}$ , where each vector  $\mathbf{I}^\beta(t)$  represents an acoustic cue (pitch, onset, ...) derived from the spectrum of the sound at time instant  $t$ . These features are extracted directly from the sound mixture, and depict a variety of possible auditory grouping cues. Details about the representation of these features and how they are extracted will be reviewed in the following section. For the time being, we need to note that each feature  $\mathbf{I}^\beta(t)$  consists, in principle, of patterns belonging to *only one* auditory object (Figure 5.4); in other words, the pre-processing stages are responsible for

segregating the sound elements into clean frames or slices belonging to a *unique* auditory object. As a concrete example,  $\mathbf{I}^\beta(t)$  may be derived from a spectrum consisting of the harmonic patterns of a specific fundamental frequency F0, which makes them features that fuse together as belonging to the same stream.

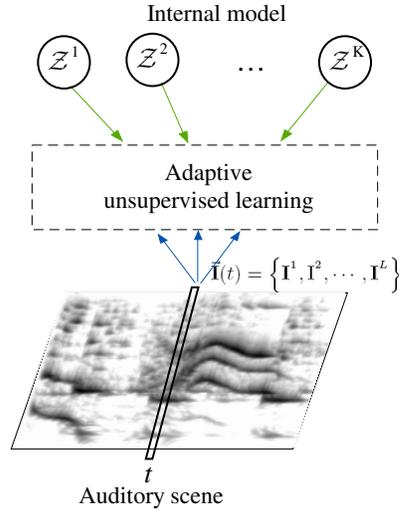


Figure 5.4: Internal world model and problem of estimation and clustering. At each time instant  $t$ , a “slice” the acoustic spectrogram is mapped into a set of feature vectors  $\vec{\mathbf{I}}(t)$ . These features are used as learning patterns to be clustered into separate clusters  $\mathcal{Z}^\alpha(t)$  representing an internal representation of the auditory scene.

Our goal is twofold: **(1)** to infer the distribution of patterns  $\{1, \dots, L\}$  into a set of  $K$  streams at each time instant  $t$  (clustering stage), and **(2)** to estimate the state of each cluster  $\alpha \in \{1, \dots, L\}$  given a newly observed input vector  $\mathbf{I}^\beta$  (estimation stage). The first goal of inferring the segregation of sound patterns to different “internally-represented” streams evokes the involvement of the brain in self-organizing the incoming sound mixtures into their corresponding perceptual units. Our second objective accredits the cortex for a role it has been attributed by many cognitive neuroscientists, that of learning and maintaining an internal model of the external world [10]. We shall address the latter estimation goal first and then take up the inference problem.

### The estimation problem

The problem of estimating the state of cluster  $\alpha$  given an observation vector  $\mathbf{I}^\beta$  can be addressed only given a formulation of the relationship between the observation and the internal model. We adopt a linear mapping between the internal state vector  $\mathcal{Z}$  and the observable sound feature  $\mathbf{I}$ , given by the equation:

$$\mathbf{I}(t) = \mathbf{A}\mathcal{Z}(t) + \nu(t) \quad (5.1)$$

where the matrix  $\mathbf{A}$  designates how the attributes of the internal model are transformed to yield a measurable sound feature  $\mathbf{I}$ . Equation 5.1 is referred to in the literature as “the measurement equation” [123], since it captures how the measured sound features  $\mathbf{I}$  of a particular perceptual stream result from a set of underlying factors  $\mathcal{Z}$  that cannot be directly observed. For simplicity, we keep this equation in general terms without specifying any superscripts to either  $\mathcal{Z}$ ,  $\mathbf{I}$  or  $\mathbf{A}$  in order to describe the general transformation from internal model representation to observation space. The real transformation, however, involves a relation between the state of one auditory object  $\alpha$  giving rise to a specific sound feature  $\beta$ . The term  $\nu(t)$  represents a noise term that accounts for any discrepancy not accounted for by the linear relationship between  $\mathcal{Z}$  and  $\mathbf{I}$ . The process  $\nu(t)$  is also called “measurement noise”, which we take to be a zero-mean Gaussian process with covariance matrix  $\Sigma$ .

A complete formulation of the estimation problem requires the inclusion of another equation that defines the dynamics governing the progression of the state vector  $\mathcal{Z}$  in time. Such time evolution is captured by the equation:

$$\mathcal{Z}(t+1) = \mathbf{B}\mathcal{Z}(t) + \eta(t) \quad (5.2)$$

The matrix  $\mathbf{B}$  is a time-invariant “state transition” matrix [123]. It represents the available knowledge of how the model’s attributes change over time, and is an appropriate parameter to invoke any constraints or expectations of the evolution of sound streams, such as temporal continuity. The “state noise” term  $\eta(t)$  is also considered to be Gaussian with zero mean and covariance matrix  $\mathbf{Q}$ . It accounts for any variability in the dynamics of  $\mathcal{Z}$  not captured by linearity.

To infer  $\mathcal{Z}$  from  $\mathbf{I}$ , we opt for an optimal vector that maximizes the model’s posterior probability. The optimization function is then defined as:

$$\begin{aligned}
\mathcal{J} &= \max P(\vec{\mathcal{Z}}|\vec{\mathbf{I}}) \\
&\stackrel{a}{=} \max \prod_{\alpha} P(\mathcal{Z}^{\alpha}|\mathbf{I}) \\
&\stackrel{b}{=} \max \prod_{\alpha} P(\mathcal{Z}^{\alpha}|\mathbf{I}^{\beta}) \\
&\stackrel{c}{=} \max \sum_{\alpha} \log P(\mathcal{Z}^{\alpha}|\mathbf{I}^{\beta}) \\
&\stackrel{d}{=} \max \sum_{\alpha} \left[ \log P(\mathbf{I}^{\beta}|\mathcal{Z}^{\alpha}) + \log P(\mathcal{Z}^{\alpha}) \right] \\
&= \min \sum_{\alpha} \left[ -\log P(\mathbf{I}^{\beta}|\mathcal{Z}^{\alpha}) - \log P(\mathcal{Z}^{\alpha}) \right]
\end{aligned} \tag{5.3}$$

where the derivation above depends on the following facts: (a) The state vectors  $\mathcal{Z}^{\alpha}$  are independent (since they represent independent streams in the environment), (b) at each instant  $t$ , the state vector  $\mathcal{Z}^{\alpha}$  depends only on a unique feature vector  $\mathbf{I}^{\beta}$  (i.e., each stream learns from only one observed input at a time, as shall be addressed in the next section), (c) the log function is a monotonically increasing function, and (d) we apply Bayes rule to transform the posterior distribution  $P(\mathcal{Z}^{\alpha}|\mathbf{I}^{\beta})$  into a likelihood and prior product, where  $P(\mathbf{I}^{\beta})$  acts as a normalization constant and can be dropped from the optimization function. The choice of  $\beta$  (in step  $b$ ) is defined by a set of constraints that we will elaborate on later.

For the time being, we assume that we have inferred the correct feature  $\mathbf{I}^\beta$ , and we need to predict the model's state  $\mathcal{Z}$ .

Given the Gaussian distribution that characterizes the measurement equation (Equation 5.1), we can reduce the optimization function  $\mathcal{J}$  to a *least-square criterion* [7]. To do so, we approximate the probability distribution of  $\mathcal{Z}$  by a Gaussian process around a mean value  $\hat{\mathcal{Z}}(t)$ , representing the current estimate of  $\mathcal{Z}$  at time  $t$ , and a covariance matrix  $\Pi$ , depicting the variability in the estimate of  $\mathcal{Z}$  in time. Consequently, the posterior probability optimization function in Equation 5.3 reduces to:

$$\mathcal{J} = \min_{\alpha} \sum_{\alpha} \left[ (\mathbf{I}^\beta - \mathbf{A}\mathcal{Z}^\alpha)^T \Sigma^{-1} (\mathbf{I}^\beta - \mathbf{A}\mathcal{Z}^\alpha) + (\bar{\mathcal{Z}}^\alpha - \mathcal{Z}^\alpha)^T \Pi^{-1} (\bar{\mathcal{Z}}^\alpha - \mathcal{Z}^\alpha) \right] \quad (5.4)$$

Minimizing this function corresponds to finding the optimal vector  $\hat{\mathcal{Z}}^\alpha$  by setting  $\frac{\partial \mathcal{J}}{\partial \mathcal{Z}^\alpha} = 0$ .

The derivation is detailed in Appendix A, section A.1, resulting in the equation

$$\hat{\mathcal{Z}}^\alpha = \bar{\mathcal{Z}}^\alpha + \mathbf{G} (\mathbf{I}^\beta - \mathbf{A}\bar{\mathcal{Z}}^\alpha) \quad (5.5)$$

where  $\mathbf{G} \triangleq \left( \mathbf{A}^T \Sigma^{-1} \mathbf{A} + \Pi^{-1} \right)^{-1} \mathbf{A}^T \Sigma^{-1}$ . The optimal solution can be formulated in words as: *New estimate = Old estimate + Gain x residual error*.

This derivation leads directly to a Kalman filter formulation [33], which offers a recursive solution to the linear filtering problem (Equations 5.1 and 5.2) under assumptions of Gaussian process and state noises. The Kalman solution is in fact a *maximum a-posteriori* estimator (MAP), which results from the optimization function in Equation 5.3. Kalman filtering is one of the best known algorithms for accurately estimating and predicting the internal state of an observed dynamical system. At each time instant, the filter updates its estimate of the model's state by improving the old estimate given the

newly observed data. The correction term is often referred to as *sensory innovation* or *sensory residual error*. It is captured by the term  $\mathbf{I}^\beta - \mathbf{A}\bar{\mathbf{Z}}^\alpha$  in Equation 5.5 and reflects the discrepancy between the predicted measurement  $\mathbf{A}\bar{\mathbf{Z}}^\alpha$  and the actual observation  $\mathbf{I}^\beta$ . If the two are in complete agreement, the residual becomes zero, and hence the *a priori* estimate  $\bar{\mathbf{Z}}^\alpha$  corresponds exactly to the predicted *a posteriori* estimate. The gain term  $\mathbf{G}$  is a weighting factor indicating how much to “trust” the new observation against the predicted *a priori* estimate. It adjusts as a function of the variance in the observations  $\Sigma$  as well estimate error covariance  $\Pi$ . By formulating the optimal solution  $\hat{\mathbf{Z}}^\alpha$  in Kalman filtering terms, we can directly apply the Kalman equations to implement a solution to the estimation problem [33, 123].

### The clustering problem

The second goal set up for our adaptive learning algorithm is that of cluster allocation, i.e., inferring which cluster (stream)  $\alpha$  “wins” the input vector  $\mathbf{I}^\beta$ . This step invokes the unsupervised learning attribute of neural processing. As mentioned earlier, competitive learning is an appealing choice for cortical learning which allows the distribution of patterns from several streams to be assigned to different clusters.

At each instant in time, a set of input features representing information about the sound mixture “compete” amongst each other, in deciding which pattern gets attributed to which class. The winning cluster for a specific pattern is the one that *matches* the best the internal state of that cluster. The key element in inferring the distribution of input patterns among the different clusters relies on a suitable choice for a similarity measure. Given the formulation of the estimation problem as a Kalman prediction scheme, an appropriate choice of distance measure is one that minimizes the sensory residual error

between the *predicted* input projected by the internal world model, and the actual *observed* sensory pattern representing the state of the environment at that particular moment in time (Equation 5.5). Hence, the optimal clustering solution for each cluster  $\alpha$  can decidedly be formulated as:

$$\mathcal{E} = \min_{\beta} \left( \mathbf{I}^{\beta} - \mathbf{A}\mathcal{Z}^{\alpha} \right) \quad (5.6)$$

which results in the following overall learning function:

$$\hat{\mathcal{Z}}^{\alpha} = \bar{\mathcal{Z}}^{\alpha} + \mathbf{G} \left( \mathbf{I}^{\arg \min_{\beta} [\mathbf{I}^{\beta} - \mathbf{A}\bar{\mathcal{Z}}^{\alpha}]} - \mathbf{A}\bar{\mathcal{Z}}^{\alpha} \right) \quad (5.7)$$

The above formulation combines the clustering problem with the optimal solution derived for the estimation problem, laying the foundation of the methodological approach for the adaptive learning component of the model. The specifics of the model architecture and implementation are going to be addressed next, and particularly the principles and parameters of the predictive component of this approach: the “internal” world model.

### 5.3.3 Model architecture

The overall architecture of the system starts with an analysis of the sound mixture through a model of peripheral auditory processing to extract an auditory spectrogram. The model used is comparable to that described in chapter 2, and given by Yang *et al.* [147]. Using this initial spectrotemporal representation of the sound input, a stage of primitive cues extraction is performed, where acoustic grouping cues are selected as described earlier in section 5.1.2. The outcome of this stage is to build a map of pitch traces and onset patterns, as will be detailed in the implementation section.

The core of the model operates along the temporal axis, one time “slice” at a time. It

examines each feature set at a given time instant through a multi-scale analysis, aimed at analyzing timbre and spectral patterns in the input. This analysis is performed following the multi-scale representation offered by the cortical model described in chapter 2.

The main component of the model is the adaptive learning module. We choose principals of cortical processing as a general scheme for implementing the estimation and clustering problems defined earlier. We define each cluster as a set of “cortical filters”; where the term cortical filter is used here to represent a filter tuned to a specific temporal modulation rate. These functions are organized in a filter-bank structure. Each one of these filters is characterized by its transfer function, which yields a set of parameters that describe its dynamics. The parameter set from the entire bank of filters define the value of system parameters  $\mathbf{A}$  and  $\mathbf{B}$  governing the estimation problem, while the filters’ internal states are captured by the variables  $\mathcal{Z}$ . Different clusters are hence constructed of various learning units, allowing each cluster to span a range of temporal scales. All clusters used in the algorithm share the same range of temporal dynamics. We shall describe in the following section how the mapping from the cortical filter-bank to the derivation of a state-space formulation is performed.

The Kalman-based algorithm (Figure 5.5) progresses in two steps: (1) a *prediction* step, which estimates the process state at time  $t$ , given previous states, and a temporal continuity constraint where the current system output is assumed not to change from one time instant to the next, and (2) a *correction* step, where this a priori estimate is corrected by the observed input. The first step is responsible for projecting the estimates forward in time to obtain *a priori* estimates for the next time instant. The second step is a feedback stage, which uses the new observation to improve the state prediction, yielding a corrected *a posteriori* estimate. This *a posteriori* estimate is then used to project forward in time,

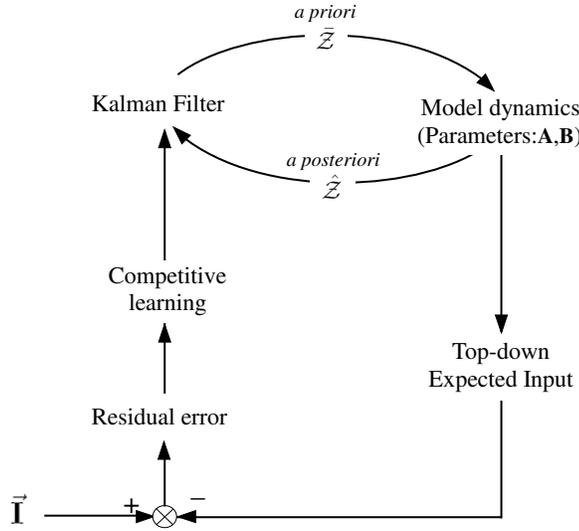


Figure 5.5: Schematic of the stream segregation model.

predicting the new *a priori* estimate for the next time step. We will describe the various stages of the model in the following implementation section.

## 5.4 Implementation of the model

### 5.4.1 Pre-processing stage

The starting point for analyzing an auditory scene is performing a spectral estimation of the sound. The signal undergoes a series of transformations converting the one-dimensional time waveform to a two-dimensional time-frequency cochlear representation. The stages of this spectral transformation follow the model of auditory periphery presented in chapter 2. The auditory spectrogram reveals the spectrotemporal patterns in the auditory scene, hence setting the ground for a spectral analysis of the various acoustic primitive features. Figure 5.6 illustrates an auditory spectrogram of an auditory scene to be analyzed.

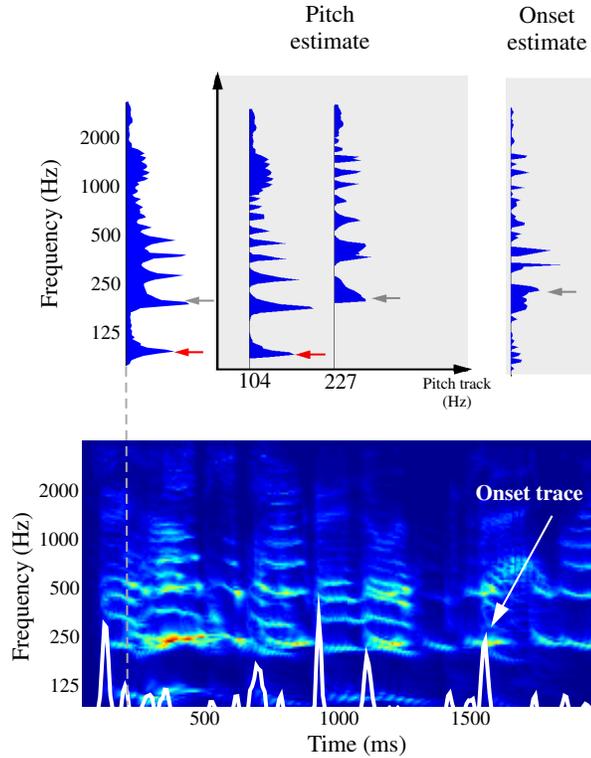


Figure 5.6: Pre-processing stages of adaptive learning model. At each time instant, a spectral “slice” in time is analyzed for any harmonicity and onset cues. The spectrogram corresponds to a mixture of a male and a female utterances. The spectrum in the upper left panel shows harmonic peaks, which are found to correspond to fundamentals of 104Hz and 227Hz. The same spectral slice also coincides with the start of syllabic segment of the female voice, and hence the onset pattern (upper rightmost panel) exhibits harmonicity relative to an F0 of 227Hz.

### Pitch estimation

The first acoustic cues to be extracted from the auditory spectrogram are pitch estimates (Figure 5.6). The goal of this stage is not to track any frequency trajectories, but simply to extract harmonic structures (if any) at every instant in time. Frequency channels that stand in harmonic relationship to each do in fact group together, and hence represent a feature vector belonging to a common sound source. Our pitch extraction algorithm is based on a template matching model, similar to that proposed by Goldstein [73]. The

model transforms the spectral representation of a signal into a distribution of pitch estimates. Incoming spectra are compared against an array of harmonic templates, which are constructed from sequences of frequency intervals around each harmonic partial of a fundamental frequency. Pitch values are determined based on the best matches to the different template patterns. The strength of pitch estimate at different F0's is judged based on the stimulus excitation patterns that fall through the sieve intervals of the nominal F0. In other words, evidence for a pitch match at F0 is based on the overall activation energy in the spectrum that matches the F0 template. A threshold is set to choose the values of F0 with the largest evidence weight at every instant in time. Depending on the number of sound sources or harmonic structures present in the original signal, a set of 1-4 F0 matches typically emerge at every time instants, which can then be used as feature cues that contribute to the learning stage.

While no definitive conclusions about the neural mechanisms of pitch have been reached yet, there is general agreement on the perceptual and acoustic attributes giving rise to the percept of pitch [94]. The model of template matching used here has been presented as one of the biologically plausible mechanisms for periodicity pitch [36]. Work by Shamma and Klein [133] suggested a biologically inspired model for the formation of harmonic templates in the early stages of the auditory system based on the phase-locking properties of cochlear filters. Their model explains how harmonic templates can emerge in the peripheral auditory system from a simple coincidence detection network operating across cochlear channels. Though quite successful in yielding proper pitch estimates of various spectral patterns, the template matching method has been criticized for its lack of robustness, and particularly in introducing additional estimates at octave or sub-harmonic intervals of the fundamental frequency. These additional estimates are not a real concern

for our scene analysis model. The learning module uses all the estimates available to find the best match to a sound source. An F0 pattern as well as 2F0 do not present conflicting information about a sound source. The learning procedure typically picks the F0 template as a closer representation of a sound stream. The 2F0 template does not bare a harmonic compatibility with a different source, and hence would not contribute as evidence in the learning procedure.

An interesting observation from using a pitch extraction scheme is that it also contributes in separating out frequency channels that are not harmonically related. While it may seem counter-intuitive for some people to think of pitch as a separating rather than grouping cue, this concept adds in fact another dimension to role of harmonicity in scene analysis. Harmonicity is in fact playing a dual role in *bringing together* frequency channels that relate to each other, but also *pulling apart* frequency channels that should not be considered as a group. Following this interpretation, pitch is a dividing force that flags apart frequency channels that should probably fall in separate streams, unless other evidence indicates otherwise. As a whole, grouping/segregating cues are indeed playing the role of exposing any differences where necessary *and* promoting fusion or grouping where necessary. This force operating in both directions is only enhancing the salience of perceptual similarities or differences between acoustic patterns represented in a multidimensional perceptual auditory domain, and hence contributing to the parsing of auditory scenes.

### **Onset estimation**

Along with pitch estimates, onset maps are very effective and robust representations of single sound sources. Onset synchrony is a particularly powerful cue for segregating acous-

tic components. We employ a simple derivative approach via temporal differentiation to boost the detection of transient energy in the signal. We then proceed to a spectral integration across frequency bands, followed by an energy threshold. Synchronous frequency channels that get activated together emerge an onset spectral segments to be used as input vectors in the learning module. Figure 5.6 illustrates the extraction of onset patterns from an auditory spectrogram. The onset trace shown overlapping with the spectrogram highlights the time instants when energy transient are detected. Those peaks are used to extract the onset pattern from the temporal-differentiated spectrogram. Such pattern, as shown in the upper rightmost panel of Figure 5.6, tends to coincide with one sound source, since it is highly unlikely that two sound sources start at exactly the same time instant.

#### 5.4.2 Multi-scale representation

Spectral shape is an effective physical correlate of the percept of timbre [32, 145]. Inspired from findings of cortical spectral analysis, we employ a multi-scale model based on a wavelet decomposition of spectra into an array of local and global spectral patterns, as introduced in Chapter 2. This spectral decomposition offers an insight into the timbre components of each of the acoustic features extracted so far. The local and global spectral shapes in the acoustic patterns are captured via a bank of spectral modulation filters tuned at different scales (0.25 – 4 cycles/octave). On the one hand, the slowest modulation scales capture the general trend in the spectrum, hence highlighting the components with broad or coarse spectral attributes, such as speech formants. On the other hand, the high-order scale coefficients describe the more dense spectral patterns corresponding to features with higher spectral density, such as harmonic peaks. Unlike cepstral analysis commonly used

in speech recognition systems, the multi-scale model operates *locally* along the tonotopic frequency axis.

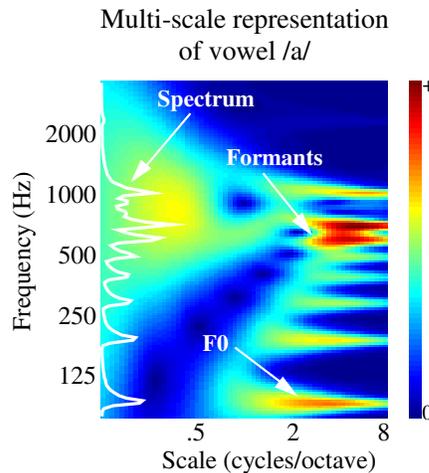


Figure 5.7: Multi-scale auditory representation of the vowel `\a\` of a male voice. The multi-scale mapping highlights various features of the original spectrum, namely the fundamental frequency  $F_0$  and its harmonic partials, as well as the formant peaks of the vowel (particularly  $F_2$  and  $F_3$ ).

Figure 5.7 shows the spectral patterns originating from the vowel `\a\` produced by a male speaker. The multi-scale representation exhibits various interesting features about the original spectrum: The location of the fundamental frequency and harmonic partials, the location of the formant frequencies. The Harmonic patterns are most notably expressed at the high spectral scales (above 2 cycles/octave). The coarse structure of spectrum, peaking at about 500-1000Hz is captured by the lower scales, which tend to exhibit the overall global patterns of a spectrum. These low scales are in fact broadly tuned filters that maintain the global shapes of sound patterns, and hence reflect the general energy patterns in the spectrum. In the results shown later in this chapter, we use spectral modulations in the range 0.125 – 4 (cycles/octave).

### 5.4.3 Cortical filtering

The temporal dynamics of the model are shaped according to a cortical multi-rate analysis similar to the one described in Chapter 2. This model is augmented by a predictive learning step as described earlier. Specifically, an input multi-scale feature (2D pattern of frequency-scale) is used as the external input pattern  $\mathbf{I}^\beta(t)$ . At the same time instant  $t$ , the model uses its past state  $\mathcal{Z}(t-1)$  to predict what the expected input should be. This prediction is performed using a temporal coherence constraint where the past output of all cortical filters is assumed to stay constant. The Kalman formulation is hence used to predict the expected input at time  $t$ . This expected input is compared with the actual observed input  $\mathbf{I}^\beta(t)$ , and the residual error between the two is used to update the current state of the filters as well as their current output  $Y(t)$ .

To be able to apply the Kalman-predictive learning technique, we have to reformulate the cortical filtering process in term of Equations 5.2 and 5.1 (state and measurement equations). This derivation is performed in three steps:

1. Writing the model dynamics in terms of a difference equation, relating the input  $\mathbf{I}$  and the output  $Y$ , via parameters  $\mathbf{A}$  and  $\mathbf{B}$ .
2. Converting the difference equation into state-space form, by defining a vector of state variables  $\mathcal{Z}$ .  $\mathcal{Z}$  represents the model's internal representation, which we choose to be the state of the delay registers in a direct form II implementation of the difference equation mentioned in step 1 [123]. The state-space form involves then two equations: (a) The process equation, representing the dynamics of the variable  $\mathcal{Z}$ , in other words, the prediction  $\bar{\mathcal{Z}}$  of the state  $\mathcal{Z}(t)$  given prior states  $\{\mathcal{Z}(t-1), \dots, \mathcal{Z}(0)\}$  (i.e., based on prior data), and (b) a measurement equation which relates the data

to the state variable  $\mathcal{Z}$ . In our case, we are using the internal model to predict our expectation from the world, and so we formulate the state-space equation to predict the sensory input  $\mathbf{I}$ .

3. From the state-space formulation, we can apply a Kalman filtering strategy. However, our ultimate goal is really to use this prediction for clustering sensory inputs into different streams. So, we proceed with the Kalman prediction, but at the update stage, we only update the states belonging to the cluster that minimizes the residual error.

Next, we elaborate on each one of these steps.

\* *Writing a difference equation:*

The dynamics of the model are governed by band-pass filters representing a multi-rate analysis of the sensory input. The impulse response of each filter is defined by a basic gamma probability density function (similar but slightly modified from the one defined in [31]):

$$h(t) = t^2 e^{-3.5t} \sin(2\pi t)$$

The impulse responses of filters at different rates is given by a dilation operation and sinusoidal interpolation of the function and its Hilbert transform:

$$\begin{aligned} h(t; \omega) &= \omega h(\omega t) \\ \Rightarrow h(t; \omega, \theta) &= h(t; \omega) \cos \theta + \hat{h}(t; \omega) \sin \theta \end{aligned} \tag{5.8}$$

where  $\hat{h}$  stands for the Hilbert transform of  $h$ . This filtering operation can be approximated by a difference linear model. We use the Steiglitz-McBride approximation [98] for finding an IIR filter with the prescribed complex-valued time domain impulse response  $h(t; \omega, \theta)$ .

The approximation leads to a difference equation of the form:

$$a_0Y(t) + a_1Y(t-1) + \dots + a_nY(t-n) = b_0X(t) + b_1X(t-1) + \dots + b_{n-1}X(t-n+1) \quad (5.9)$$

where the vector of coefficients  $\{a_2, \dots, a_n\}$  and  $\{b_2, \dots, b_n\}$  represent the filter parameters (nominator and denominator coefficients), and can be adjusted to appropriate sizes by letting some entries take a zeros value. Writing the parameter vector  $\vec{a}$  and  $\vec{b}$  to the same length simplifies the transition from a difference equation to a state-space form, as we shall see next.

\* *Conversion to state-space form:*

State-space methods are standard system formulations that modern control theory resorts to when dealing with system dynamics and time-series analysis. Any  $n^{th}$  order difference equation can be represented by a first-order vector equation with a state vector of  $n$  elements [123]. We want to transform the formulation of equation 5.9 into a state-space form.

We perform a little twist in the formulation by actually using the past output  $\{Y(t-1), Y(t-2), \dots\}$  to predict the current input  $X(t)$ . This does not change the actual procedure of writing a state-space equation, but only affects what the actual values of parameters  $\mathbf{A}$  and  $\mathbf{B}$  are. It also allows to formalize our temporal constraint by also setting  $Y(t) = Y(t)$ . Hence, the problem is now defined as: given  $\{Y(t-1), Y(t-1), \dots\}$  and  $\{X(t-1), X(t-2), \dots\}$ , what is the value of  $X(t)$ ? In this sense, the right-hand side of equation 5.9 represents the measurable process. This derivation of a state-space equation is quite straight-forward, and is presented in detail in section A.2 of Appendix A. The final representation of the difference equation is now given by:

$$\vec{Z}(t) = \mathbf{B}\vec{Z}(t-1) + \mathbf{C}Y(t) \quad (5.10)$$

$$X(t) = \mathbf{A}\vec{Z}(t) \quad (5.11)$$

where,

$$\mathbf{A} \triangleq [a_0 \cdots a_{n-1}] \quad \text{and} \quad \mathbf{B} \triangleq \begin{bmatrix} -b_1 & \cdots & -b_{n-1} & -b_n \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{C} \triangleq \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

\* *Kalman filter:*

The Kalman theory is a well-established scheme for data estimation. By formulating the cortical dynamics based on Equation 5.11, we can directly apply Kalman equations in implementing the prediction module. The vectors  $Z(t)$  represent auxiliary parameters (or hidden states) of the cortical model, and hence maintain an internal “image” of the data in each cluster, as input observations are coming in. This cortical internal representation can in effect be thought of in terms of a perceptual organization of sound patterns, which is later translated into actual streams or auditory objects.

#### 5.4.4 Adaptive learning

The step of predictive learning is performed based on the dynamics schematized in Figure 5.5. The learning happens as the cycle between the model prediction and correction (based on observed inputs) guides the revision and alteration of cluster representations, and allows for a more accurate clustering function. Most simulations shown here limit the number of

cluster  $K$  to 2. These two clusters are in effect symbolizing auditory streams pertaining to the foreground vs. background in an auditory scene. This distinction is not necessary for the actual implementation of the algorithm, but it leads to a more intuitive understanding of the response patterns of the cortical filters. Following this interpretation, one can think of one stream/cluster as the object/sound attended to, while everything else is falling in the background, and hence in the opposite stream.

As interesting note concerning the behavior of the rate filters in the cortical model is the fact that they are guided by their internal dynamics (as determined by their individual impulse responses). Each rate filter “learns” according to the range of temporal modulations and orientation it is tuned to (the orientation selectivity is obtained in conjunction between the filter’s response and the input scale representation). The time constants of these filters in effect act as a “memory” component, whereby the neuron learns at a certain rate. If no valid input is coming in, the neurons stop learning; which is manifested by the relaxation of their outputs as they tend to drop toward zero, indicating a state of “forgetting” of those particular filters.

## 5.5 Results

In the introduction to this chapter, we described two aspects pertaining to auditory scene analysis; namely, auditory streaming and particularly of alternating sounds, and segregation of concurrent sound sources. Using our model, we investigate aspects associated with both phenomena, and focus on extending our understanding of how each module in our adaptive scheme interacts with the others in achieving the specified learning goal. Simulating various sound organization effects is both beneficial in terms of testing the capabilities of the model, as well studying the actual response patterns obtained from the cortical model. Inspired from neural evidence in the auditory system, the model is in fact a reflection of how we expect the stream formation to be happening in biology, of course formulated in neural vocabulary, rather than abstract modelling terminology.

### 5.5.1 Streaming effects

Streaming effects are very commonly tested using cycles of repeated stimuli. Whether for psychoacoustic testing or physiological experimentation, cycling stimuli are chosen for their relative ease of manipulation. They also produce a very stable perceptual manifestation after a certain number of repetitions. It is now well established that streaming is a phenomenon that builds up over time before the perceptual manifestation of the streams is well established for the listener. This scheme of cycling stimulus patterns is well suited to be employed in our learning model, as any neural network structure does require to accumulate certain evidence from its inputs before any learning or computational judgements can be made. The duration required for the buildup of a stable representation varies depending on the complexity of the sound patterns and the duration of the individual sound elements constructing the entire sequence.

Here, we describe a series of simulations of commonly known streaming paradigms which illustrate various underlying principles of sound organization. Most these paradigms have been described in a collection of demonstrations assembled by Bregman and Ahad [22].

### **Alternating tone sequence**

Alternating tones are the most classic paradigm used for studying auditory streaming. The paradigm consists of simply repeating a tone “A” and tone “B” in a sequence ABAB... By varying the frequency separation between the tones, as well as their presentation rate (i.e. the duration of each tone), listeners are asked whether they hear one or two streams [140]. As the tone frequency separation increases and the rate of presentation becomes faster, the subjects report hearing two continuous streams, one going at a lower frequency and one at a higher frequency. Additional factors are also important in determining the subjects’s perception of one or two streams, including the listeners’ attentional readiness while performing the task, and the overall presentation time of the sequence.

The schematic in Figure 5.8 shows results obtained from our adaptive learning model. This schematic is organized from the bottom to the top to signify the information flow from the input coming signal to higher-level organization of sound into perceptual streams. In this case, the algorithm is presented with a sound sequence whose spectrogram is shown in the lowermost panel of the figure. In this particular simulation, the tones are chosen to be 500Hz and 2000Hz, each 125 msec, i.e. repeating 4 times per second. The patterns go first through a multi-scale analysis. The figure shows 4 outputs of the spectral analysis at 4 different time instants as the analysis progresses in time from tone A to tone B. The middle panels show intermediary time instants as tone A fades away, and tone B

begins to rise up. Next follows the cycle of adaptive learning via Kalman estimation and competitive learning. The outcome is reflected in the responses of different cortical rate filters after the presentation of the sequence. An interesting observation is that the filters tuned at 4Hz show a continuous pattern of activation since they are tuned to presentation rate of each sequence A-A-A... or B-B-B..., repeating at a rate of 4Hz. The cortical outputs are in effect responses to sequences A-A-A... (left blue cluster in Figure 5.8), and B-B-B... in the other cluster. The topmost two panel represent a schematic of what the perceived streams actually aggregate to. These patterns are obtained by simply summing the outputs of the different rate filters in each cluster. The fact that certain cortical filters are tuned to the presentation rate of the each tones A or B is very critical in explaining the perception of a continuous stream despite the fact that each acoustic stream by itself is in reality non-continuous. The energy response of a 4Hz filter to a 4Hz cycling sequence is in fact a continuous pattern. This fact directly relates the dynamics of streaming to the low-pass nature of cortical responses, which is a manifestation of the loss of cortical phase-locking and synaptic depression (as discussed in chapter 3).

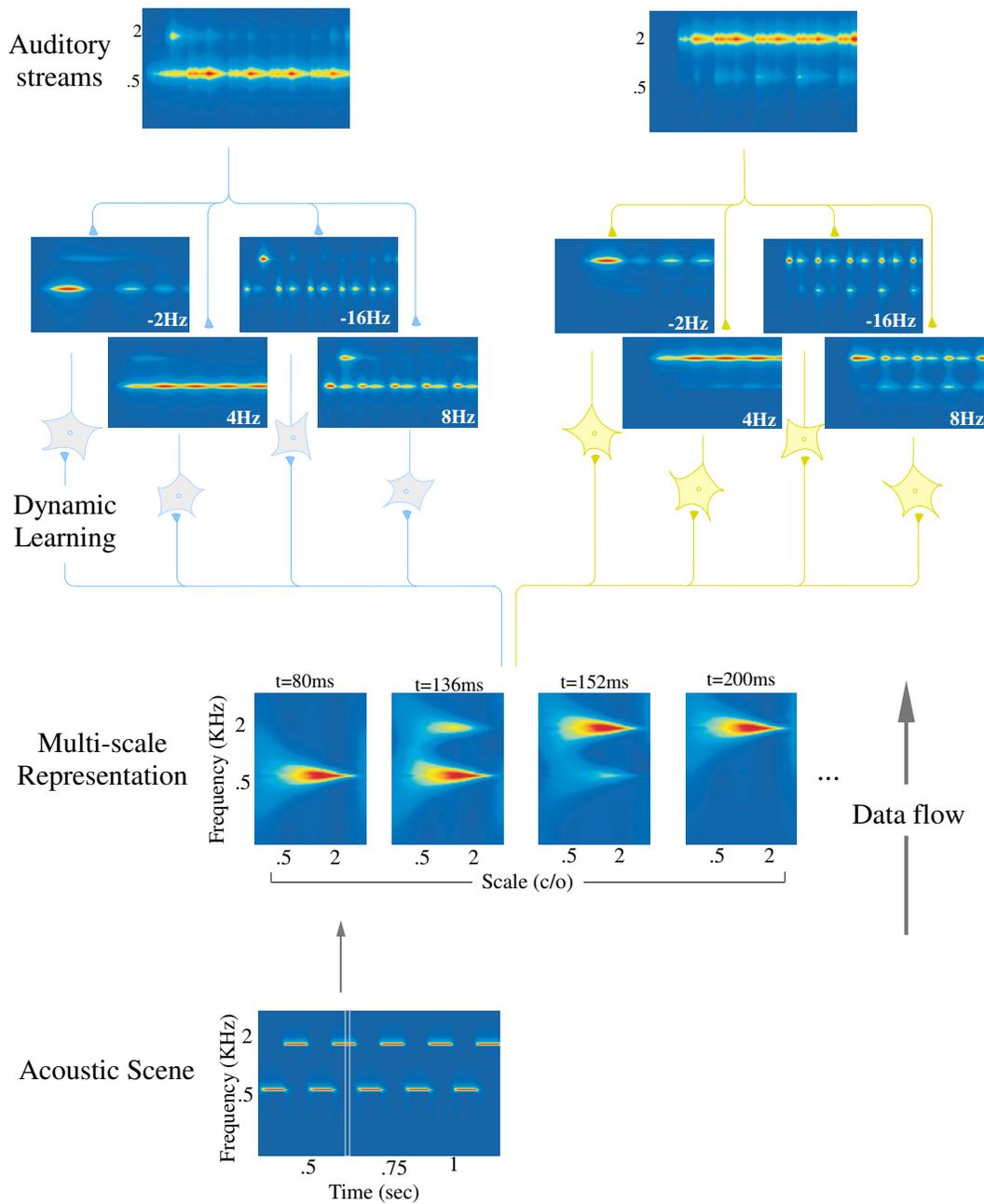


Figure 5.8: Streaming of alternating tones. The data flow in this schematic goes from top to bottom (as shown by the “data flow” arrow) indicating the bottom-up process involved in the analysis of the sound mixture. The details of this schematic are in the text. The patterns shown here correspond to the magnitude responses of the complex-valued representation in the cortical model. The spectrograms representing the auditory streams in the two topmost panels of the figure are obtained as the sum of the outputs of all rate filters within each cluster.

## Alternating ripple sequence

A similar experiment can be performed by using alternating ripple sounds. We use static ripples, in this case 0.5 cycle/octave and 2 cycle/octave. At each cycle of the alternating sequence, a new instance of a white noise carrier is generated, and modulated spectrally at either 0.5 or 2 cycles/octave (A or B sound). As we construct the entire alternating ripple sequence ABAB..., the fine spectral structure of any two sounds never repeats. This additional spin is aimed at varying any possible cues that could lead to a grouping of ripple elements, and override the role of spectral modulation patterns. The fact that carrier information is unique at each instance of either sound allows us to avoid the reliance of principles such as peripheral channelling, which would take advantage of any regularity in the frequency cues of the sound carriers in order to stream different ripples together. Using a multi-scale representation of each sound pattern, the model successfully separates these two noise patterns along a “spectral-modulation” dimension. Examples of multi-scale ripple patterns are shown in the middle panel of Figure 5.9. The leftmost panel shows a specific excitation pattern in response to a time slice coinciding with a 2 cycles/octave sound. As the sounds change to 0.5 cycles/octave, we can notice a shift in the excitation pattern (second multi-scale representation panel in Figure 5.9, at  $t = 288ms$ ). Then, the pattern completely moves toward 0.5 cycles/octave (right panels). The model successfully organizes these alternating patterns into two clusters, shown in the top panels of Figure fig:ABripples. Their spectrographic representations (in these two panels) appears a bit smeared due to the temporally-extended response structure of the cortical filters at low rates ( $\pm 2Hz$ ,  $\pm 4Hz$ ). However, they illustrate the model’s capability in organizing these two sound patterns which share a common spectral format.

Informal tests were also performed to judge listener’s ability to stream ripple sounds

with different spectral patterns. Such patterns proved to be much harder to stream when the carrier noise was randomized, but certain subjects were still able to report hearing a streaming effect when the separation between the spectral patterns of ripple A and B was large enough. This effect might be explained by prior training and acoustical or musical background of the subjects, making it easier for some to organize these sounds at a higher-level dimension (scale axis) without relying on any other dimensions.

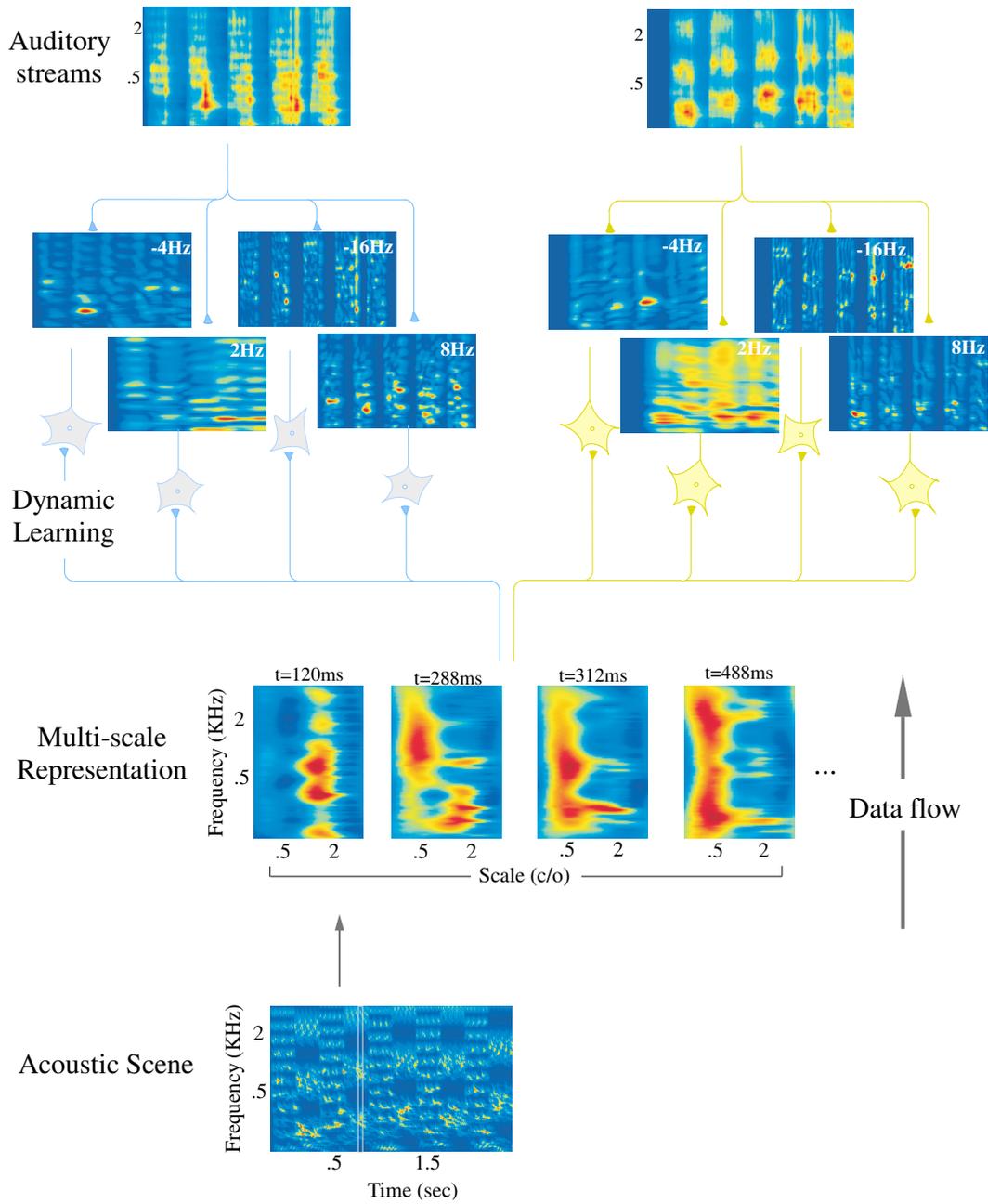


Figure 5.9: Streaming of alternating 0.5 and 2 cycles/octave ripples.

## Alternating tone cycles

Next, we focus on additional simulations collected by Bregman and Ahad [22] as a set of most common, and probably most useful, paradigms in studying organization of sound in the auditory system from a streaming perspective. We choose from this series of demonstrations those which can be implemented in our model. The first simulation is quite analogous to the alternating tone sequence ABAB.... The paradigm corresponds to psychoacoustic tests run by Bregman ([22], demonstration 1) as shown in the lower rightmost panel of Figure 5.10. The schematic represents a sequence of tone cycles alternating between high tones (H1,H2,H3) and low tones (L1,L2,L3). The actual cycle is constructed by the sequence H1,L1,H2,L2,H3,L3,... The technical details of these sound correspond to those set by Bregman [22]: The tone frequencies for the high sequence were 2500, 2000 and 1600 Hz, and for the low sequence 350, 430 and 550 Hz. These frequency ranges are well separated along the tonotopic axis. The spectral patterns shown in the middle panels of the Figure 5.10 indicate that the multi-scale representations used as input for the learning module are very well disjoint in frequency. This separation is reflected in the output of the adaptive learning algorithm, shown in the top most panels. It demonstrates how successfully we can perceive two well segregated sound patterns: a high frequency 3-tone melody and a low frequency sequence. This simulation is a direct test of the principle of *sequential integration*. The tones in the low (or high) frequency region fall in the same cluster, because the acoustic features from pattern L1 and L2 appear to be similar (by virtue of frequency proximity), and quite dissociated from the other “competing” patterns H1, H2, H3. This perception is only maintained as long as the sequences are repeated at a relatively fast rate, guaranteeing that the dynamics of the cortical model are commensurate with the presentation rate.

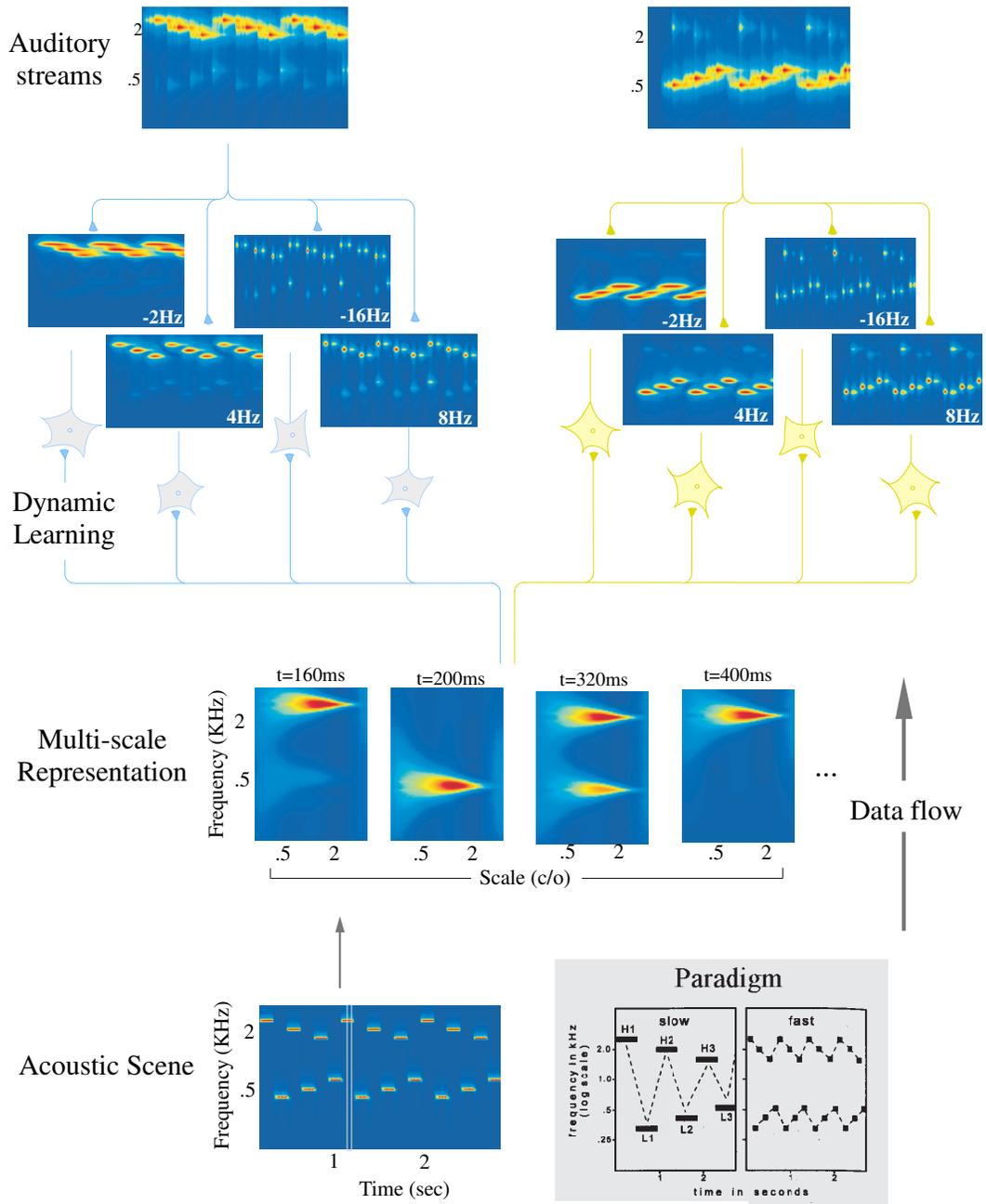


Figure 5.10: Streaming in a cycle of 6 tones. The paradigm shown in the lower right corner of this schematic corresponds to demonstration 1 in Bregman's demonstrations CD [22]

## Alternating vowels

Vowels are preferred test signals in many experiments dealing with stream segregation. They are easy to synthesize and manipulate, with different fundamental frequencies, formant structures, and excitation modes. The variability in their spectral shape is an interesting feature to explore stream segregation. We perform a simulation analogous to that carried out by Bregman ([22], demonstration 11). Instead of employing synthesized vowels as used by Bregman, we use an alternating sequence of two natural vowels /e/ and /ə/ produced by the same male speaker. The spectrogram of the sequence is shown in the lower panel of Figure 5.11. Even though they are produced at the same pitch (same speaker), the two vowels exhibit very distinct spectral shapes (different formant positions and relative intensities). The divergence between the two patterns promote streaming effects, hence contributing to their separation into two separate streams, shown in the uppermost panel of Figure 5.11.

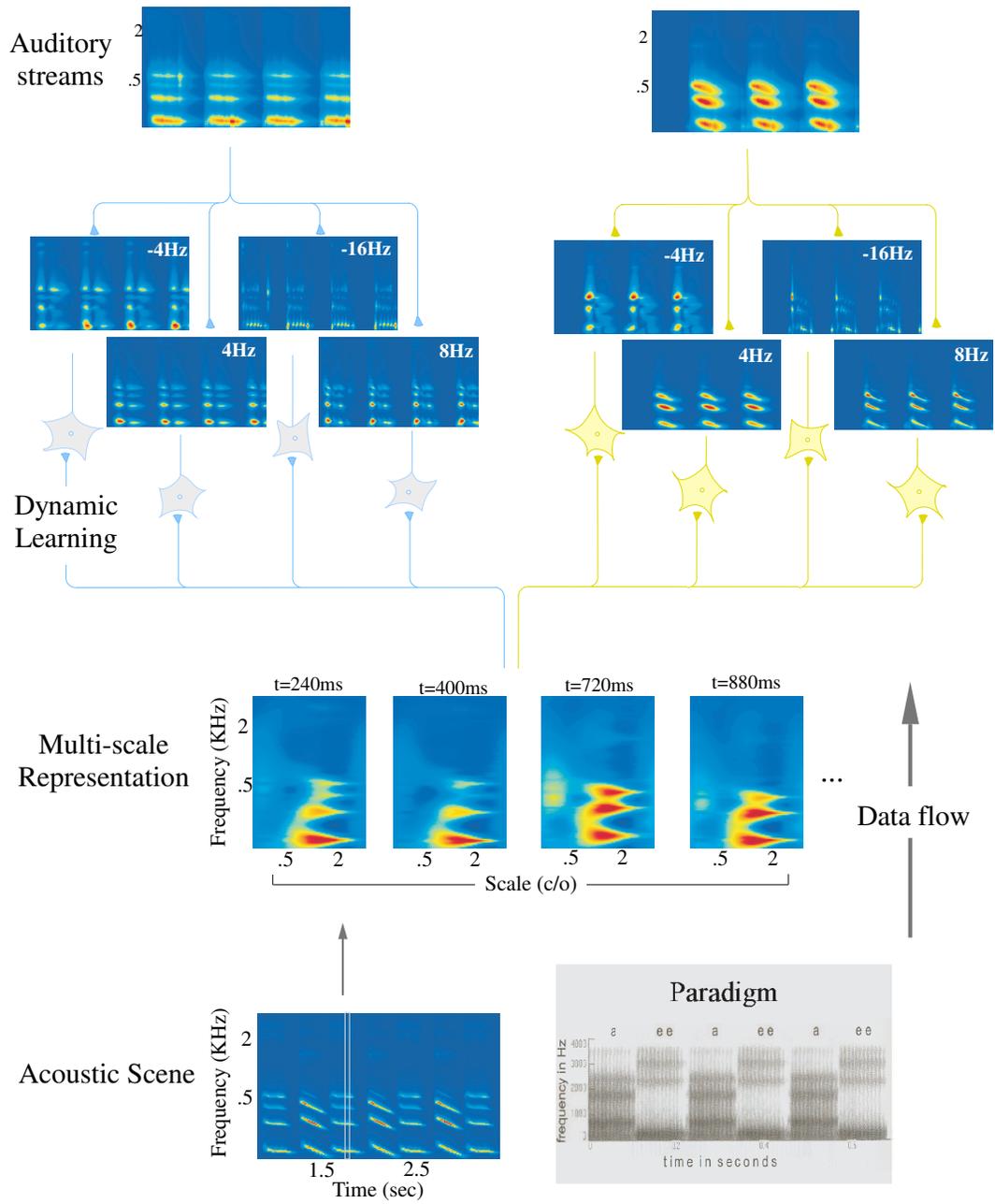


Figure 5.11: Streaming of alternating vowels. The paradigm shown in the lower right corner of this schematic corresponds to demonstration 11 in Bregman's demonstrations CD [22]

## Capturing interference tones

The involvement of attention in grouping mechanisms is of particular interest for any study of sound perception. A very powerful principle that can be easily tested in humans is orienting subjects to different streams and judging their ability to focus on certain aspects of the attended stream, while interfering patterns are played in the background. An illustration of such task is given in the paradigm of Figure 5.12, and the technical implementation details follow those of Bregman [22]. The subjects are initially presented a tone sweep AB or BA and asked to judge whether it is upward or downward going. Such judgement is very easy and quite straightforward. When surrounding the AB/BA sequence with an interfering tone X (leading to XABX or XBAX), it becomes quite hard to judge the order of the sequence AB, due to the presence of the bracketing tones. Such effect has been explained by Bregman as due to a loss of prominence of the pattern AB (or BA) since a more higher-level percept emerges due to the four-tone structure (XABX or XBAX). The effect of the surrounding interference can be almost nullified by extending the sequence X in time prior to the sequence of interest. The sound would now consist of a pattern XXXXXXABXX (Figure 5.12). The presence of a preceding sequence of X tones leads to the emergence of a separate stream made up of the “X” pattern, and hence when sound AB or BA is played, it falls into a different cluster, facilitating its release from the surrounding tones. Such release allows it to partially restore its original saliency, hence allowing us to get a better judgment of the AB-BA sequence.

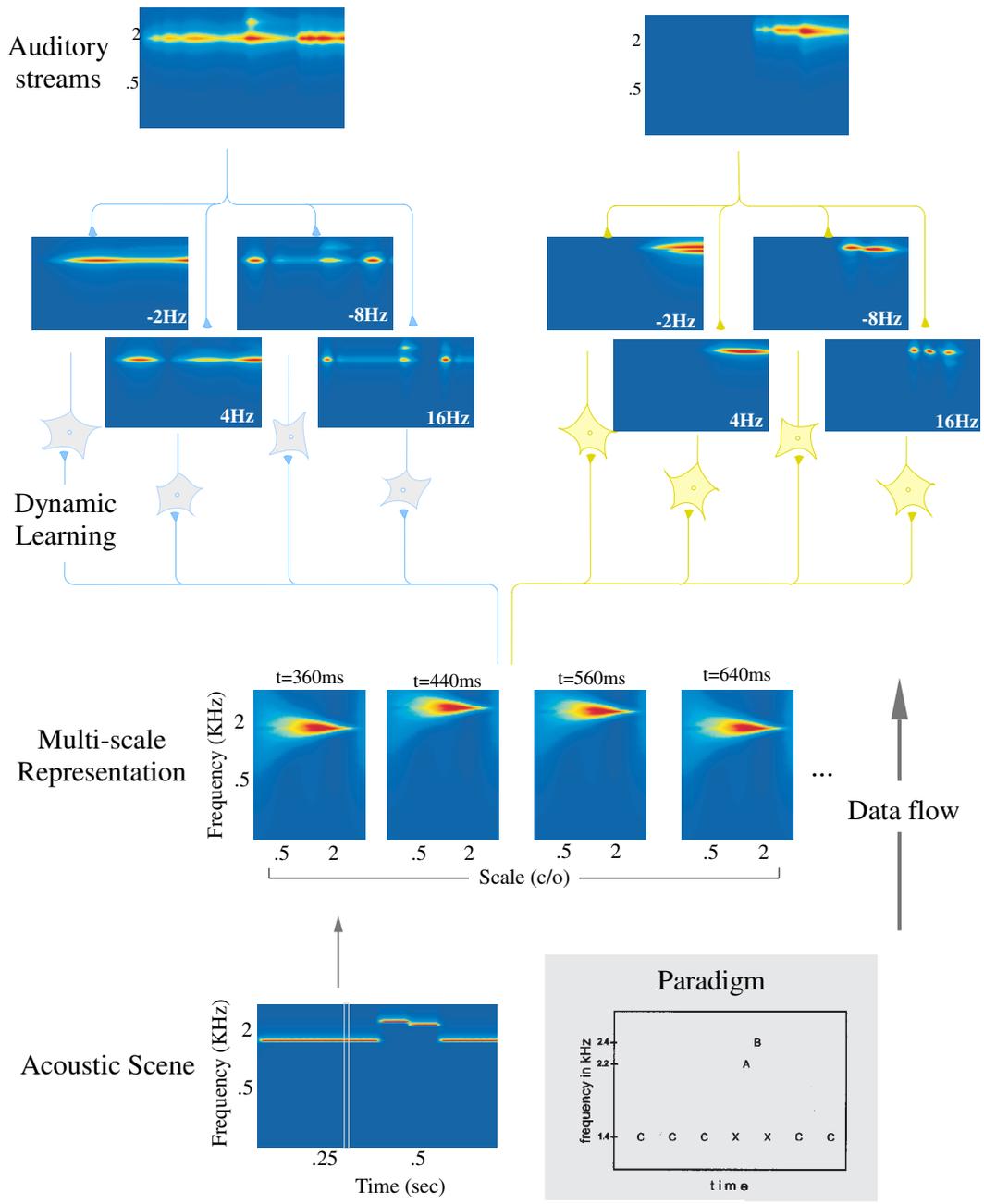


Figure 5.12: Segregation by capturing interfering tones. The paradigm shown in the lower right corner of this schematic corresponds to demonstration 16 in Bregman's demonstrations CD [22]

## Crossing trajectories

The theory of *crossing trajectories* is one that tests the segregation of rising and falling tone sequences, which overlap at a certain point in time. It is interesting to investigate whether the auditory system would actually group the rising and falling tone patterns together or not. Listeners report hearing a bouncing pattern when the sound elements are individual tones, and it is very hard for subjects to follow an entire rising or falling sequence. Such effect is illustrated in Figure 5.13, with technical details of tone frequencies and duration in agreement with Bregman's paradigm. The outcome of the learning model is quite interesting, and constitutes a direct manifestation of the bouncing effect described by Bregman. The best explanation put forward for this effect is one that favors grouping of tones belonging to a similar frequency region together.

The next test - related to this simulation - is one which tests the crossing trajectories in the case of a harmonic vs. single frequency patterns. Now, the rising pattern consists of a harmonic sequence (1<sup>st</sup>, 2<sup>nd</sup>, 4<sup>th</sup> and 8<sup>th</sup>) harmonic of a sequence with F0 at the value of the tone in the original rising pattern. By adding harmonicity to the mix, we present the system with clues suggesting that the rising pattern should be segregated into a distinct cluster, as it shares common features of periodicity. This spectral regularity influences the organization of the patterns into their corresponding streams, leading the falling sequence to now group into a distinct cluster (Figure 5.14).

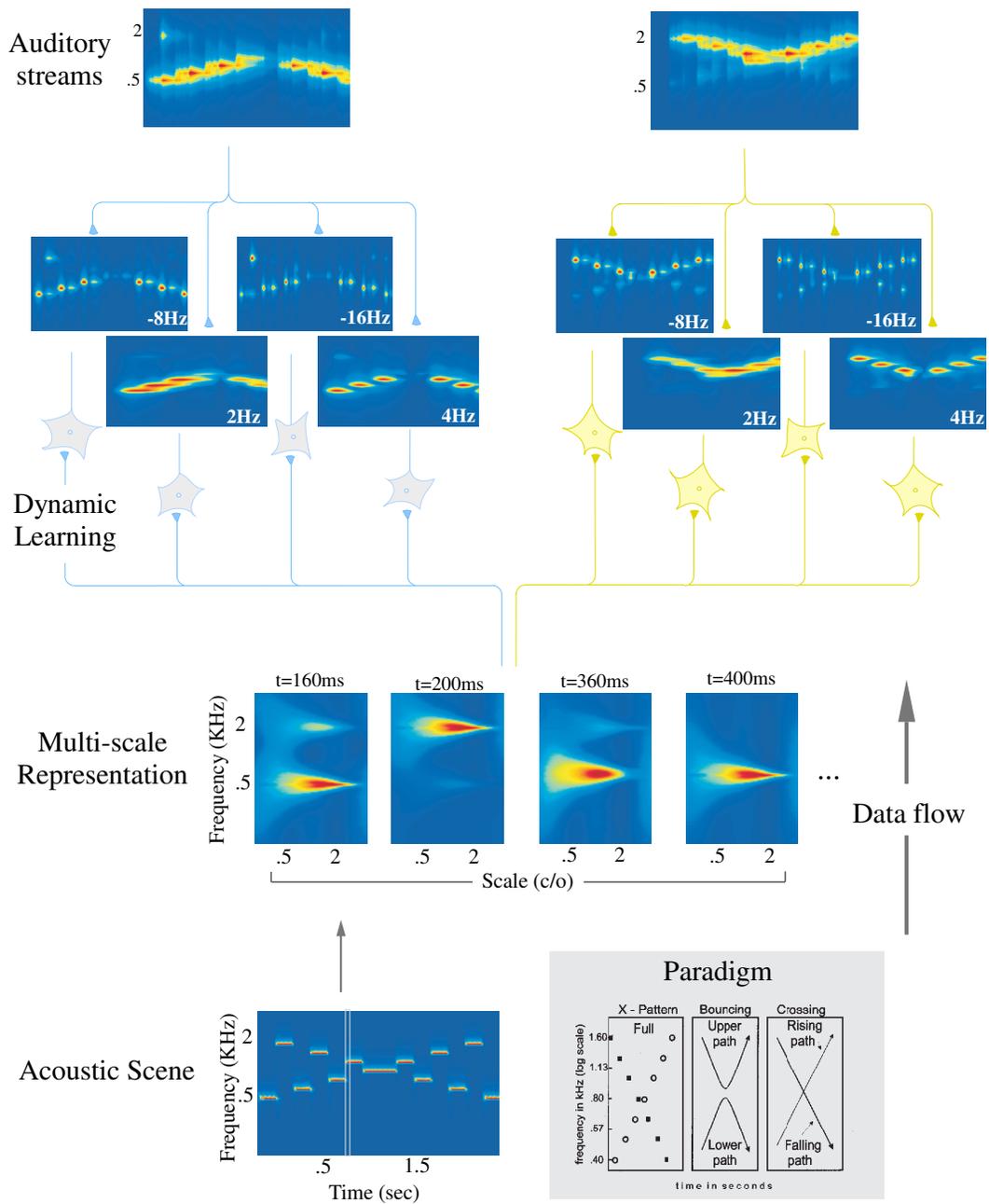


Figure 5.13: Segregation of crossing-trajectories. The paradigm shown in the lower right corner of this schematic corresponds to demonstration 17 in Bregman's demonstrations CD [22]

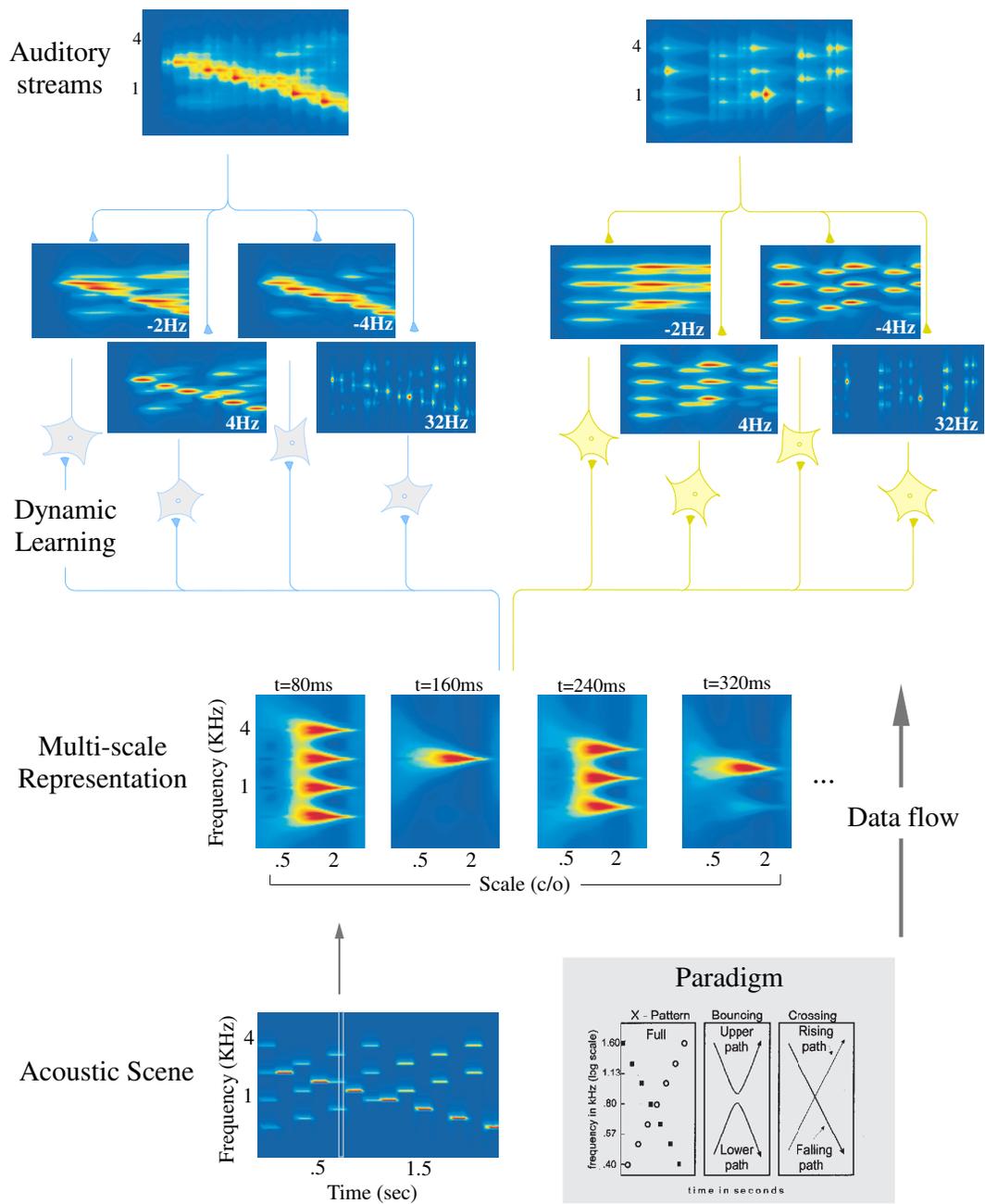


Figure 5.14: Segregation of crossing-trajectories (2). This simulation shows the second demonstration of the paradigm, when the rising sequence consists of harmonic complexes while the falling sequence is a frequency decreasing tone.

## Sine-wave speech

Language is certainly a powerful grouping feature that is hard to capture in terms of simple acoustic cues, following the principles described earlier in this chapter. Nonetheless, we are able to put acoustic elements together that sound to us like speech, even if we do not understand the language. Sine-wave speech is a special case of sound signals designed to lie at the end of a continuum representing speech patterns [125]. It consists of a number of sinusoidal waves modelling the formant tracks of speech sentences. The spectrogram in bottom panel of Figure 5.15 shows an example of such sentence. Most listeners first hear a sine-wave signal as a series of tones, chirps, and blips, with no particular hint to any particular meaning or language reference. After prompting, listeners are able to recognize the words in the sentence, and start hearing these utterances as speech, despite the very poor quality of the sound. In the context of our model, we claim that one reason why sine-wave speech originally sounds like a collection of tones and blips is simply because each sine wave is clustered into a separate stream, hence making it difficult for the auditory system to integrate information across streams to be able to recognize the linguistic meaning of the sentence. To test this hypothesis, we ran our model on a sine-wave utterance, but using three clusters to test whether the three sine waveforms would separate into the three different clusters. As expected, Figure 5.15 shows the outcome of such simulation, and confirms our prediction of separating the three constituting energy waveforms into different clusters, and hence separate streams.

In Figure 5.16, we explore an interesting extension to this idea of clustering the separate waveform into separate clusters, and hence leading to difficulty in originally recognizing the signal as speech, unless prompted. The test consists of giving the system additional hints to group the three sinusoidal trajectories together. To do so, we introduce

silences into the sentence, occurring at a rate of 20-50Hz. Interestingly, sinewave speech sounds more intelligible when segments of silence are interleaved with the original signal. The silence portions introduce common-onset cues giving evidence that the three sinusoidal waves should be integrated together in the same perceptual stream. Figure 5.16 shows an illustration of such simulation.

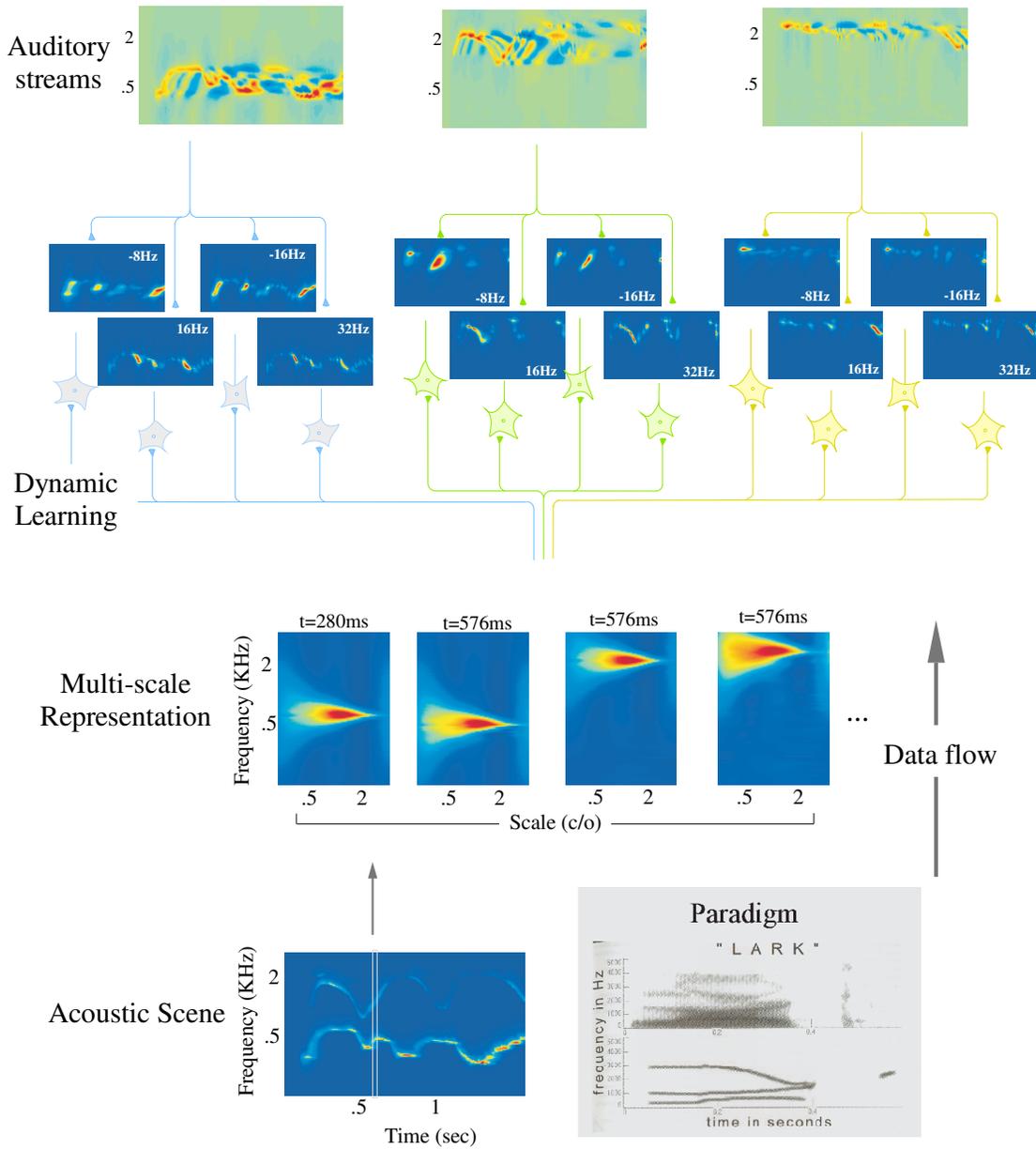


Figure 5.15: Sine-wave speech. The paradigm shown in the lower right corner of this schematic corresponds to demonstration 23 in Bregman's demonstrations CD [22]. The uppermost panels (representing the auditory streams) are now displaying the real-part of the complex-valued output of the cortical model. We choose the real-part as opposed to the magnitude in this case to better illustrate the patterns in response to fast changing patterns in sine-wave speech.

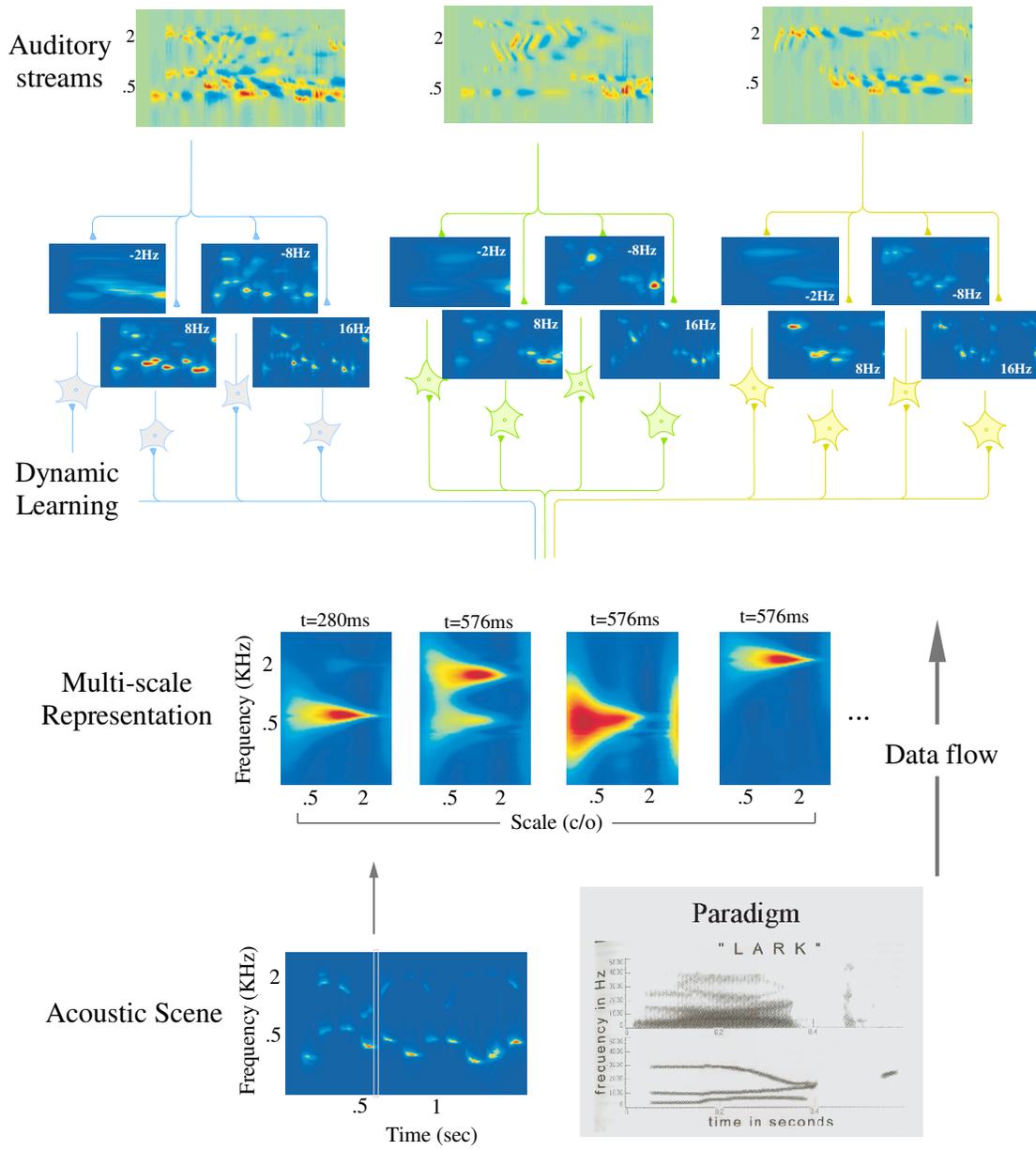


Figure 5.16: Sine-wave speech. The paradigm shown in the lower right corner of this schematic corresponds to demonstration 23 in Bregman's demonstrations CD [22]. Similarly to Figure 5.15, we again display the output of the cortical model in terms of its real-part, to better exhibit the fast patterns in the response.

## Tone in a mixture

In a final simulation of sound organization principles, we explore a paradigm which involves a different structure of auditory segregation: simultaneous grouping. The paradigm shown in Figure 5.17 illustrates a construction where different cues are put in competition with each other. The paradigm consists of a cycle of an A tone, followed by a tone complex B,C. The occurrence of tones B and C together delivers a strong onset cue, hinting that these two elements should be grouped together. In the initial simulation where tone A is played at a different frequency (1800Hz) than the complex (B: 650Hz and C: 300Hz), the streams formed by the learning algorithm segregate the scene into a stream of A sounds, and a distinct stream of BC complexes (Figure 5.17).

In a second simulation, tone A is shifted in frequency to match that of tone B in the complex BC (right panel in paradigm schematic). As the frequency separation between A and B is now zero, the system has to decide whether to group B with the already existing cluster A, or whether to use the evidence of common onset in the complex BC to cluster both tones B and C into a common new stream. Simulation results for this case are shown in figure 5.18, and do actually match our perception of these sounds. Bregman explains this result by a principle he calls “old-plus-new heuristic” [21]. This principle is proposed as an organizational scheme used by the auditory system for sound organization, and is described as: *“If a spectrum becomes suddenly more complex or more intense, the auditory system tries to interpret this as a continuing old sound joined by a new one that supplies the additional acoustic stimulation”*. In the context of the second simulation, we can interpret the results as tone A being the “old” evidence, and the complex BC supplying continuing “old” evidence (tone B), and new evidence for a new sound (C); hence the organization of these elements into an A-B-A-B.. stream versus a C-C-C-... stream.

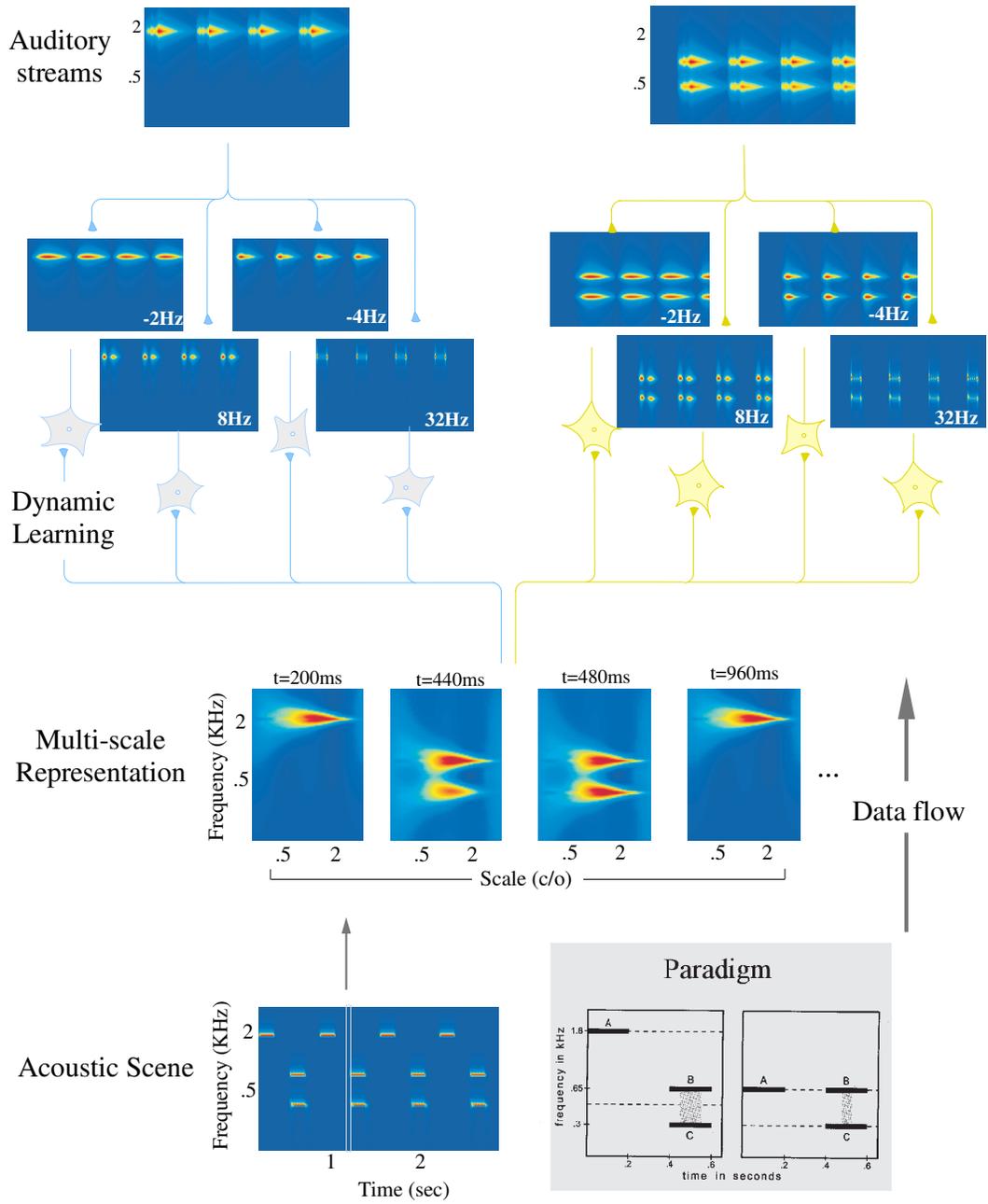


Figure 5.17: Capturing tone in mixture. The paradigm shown in the lower right corner of this schematic corresponds to demonstration 25 in Bregman's demonstrations CD [22]

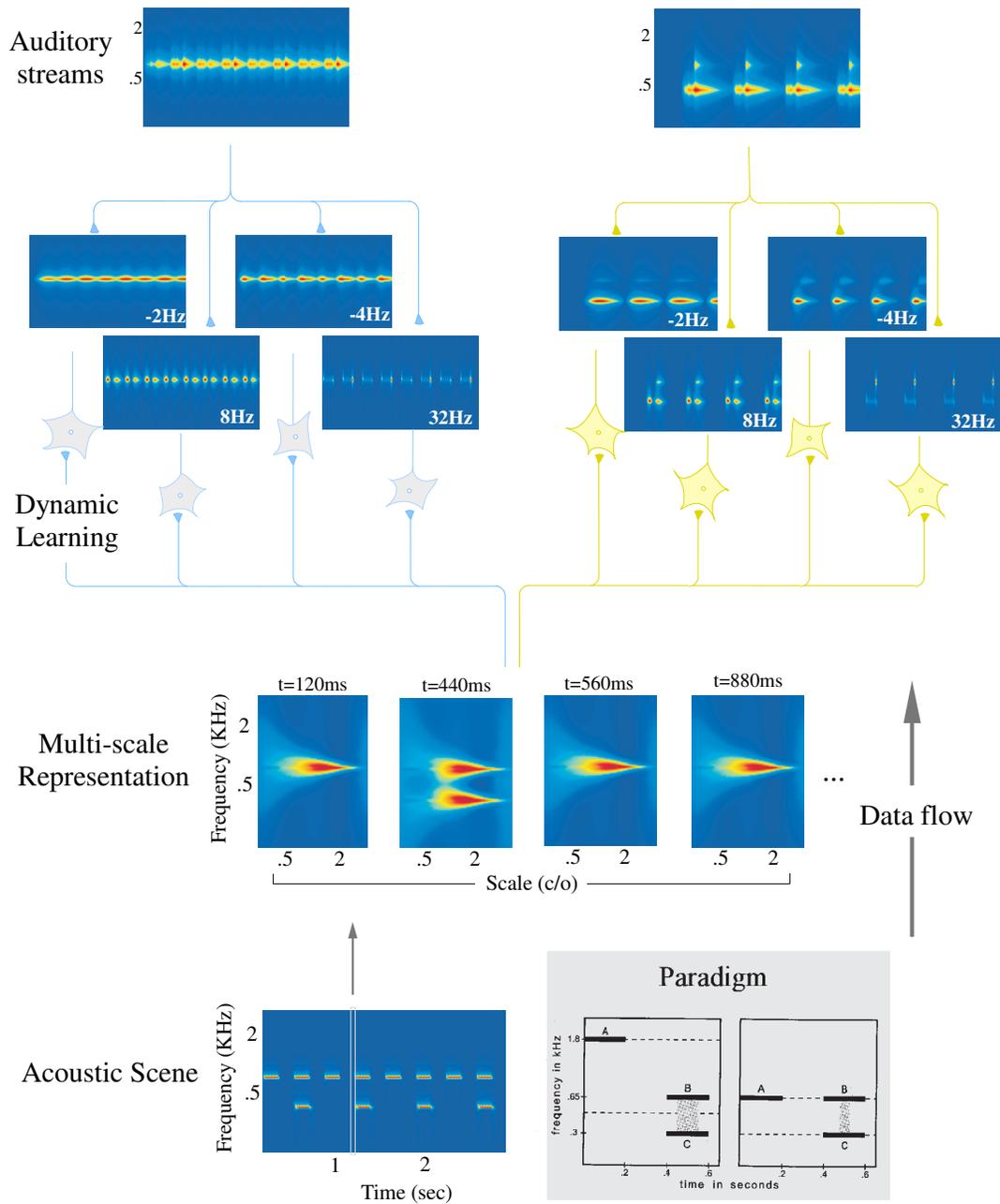


Figure 5.18: Capturing tone in mixture (2). The paradigm shown in the lower right corner of this schematic corresponds to demonstration 25 in Bregman's demonstrations CD [22]

### 5.5.2 Speech segregation

A second component of our implementation of the adaptive learning model is to test its applicability to speech separation under realistic conditions. We perform two main evaluations of the model’s performance by either using original sentences, or combining them into a sound mixture.

An important question pertaining to the evaluation method of CASA systems has been raised by most researchers working in this field [37]. The most commonly adopted approach is for each system to maintain a separate module to allow the reconstruction of sound waveforms from the segregated signals, and hence be able to run listening tests. Some studies however have supported the idea of integrating CASA systems with automatic speech recognition (ASR systems, and hence bypassing the need for a complete reconstruction of the sound waveform. Retaining a higher-level stream representation interpretable by the ASR system suffices to evaluate the outcome of the speaker separation model. These evaluation techniques are however not always easily applicable, and particularly using our model. We know of no ASR system currently available that can easily interface with our cortical representation. Moreover, the difficulty in reconstructing the acoustic waveform of the learned streams stems from the model’s extensive computational complexity. The model results in 4-dimensional complex-valued representations (time-frequency-scale-cortical rate) per cluster, which makes it very difficult to implement a re-synthesis module for the acoustic waveform. We hence use a correlation measure (normalized correlation coefficient) to evaluate the correspondence between learned patterns, and original or expected ones.

## Speech segregation using original utterances

In order to investigate how well the system separates sound pattern belonging to different speakers, we test our model using a pair of sentences from two different speakers. These two sentences are analyzed *separately* through our pre-processing stage, to basically map the sound patterns into a multi-scale representation. The features extracted from both speakers are then combined in an array of sound patterns, with no reference to which speaker they belong to. They are then clustered using the adaptive learning model. The only evidence that can indeed differentiate the features of a same speaker is the regularity in the patterns themselves, in the absence of any other labelling of its source. These patterns are then given as input to the adaptive learning model.

We run four separate set of tests using a different pair of: male-female speakers, male-male speakers, female-female speakers. A fourth test is run with a female speaker against a male speaker whose sentence has been modified so that the pitch now matches the female range, but altering the spectral ratios of his formant energies. Each one of these tests is performed with 100 different pairs of different male and female speakers and utterances from the TIMIT database [1], where sentences range between 2 and 4 seconds long. The results are shown in Figure 5.19, and are described as follows: We correlate the output of the cortical model for one cluster with the original “clean” sentence, and obtain a correlation coefficient between the two ( $\rho_l$ , or learning correlation). We also correlate the model’s responses to the two original sentences, to obtain a baseline of how well separated the original sentences are, and hence how well we can expect our model to achieve if it were to perform perfectly. We call this correlation  $\rho_b$ , or baseline correlation. Finally, we correlate the learned utterance not with its own original sentence but the other “competing” utterance, to assess the level of confusion between the patterns of the learned

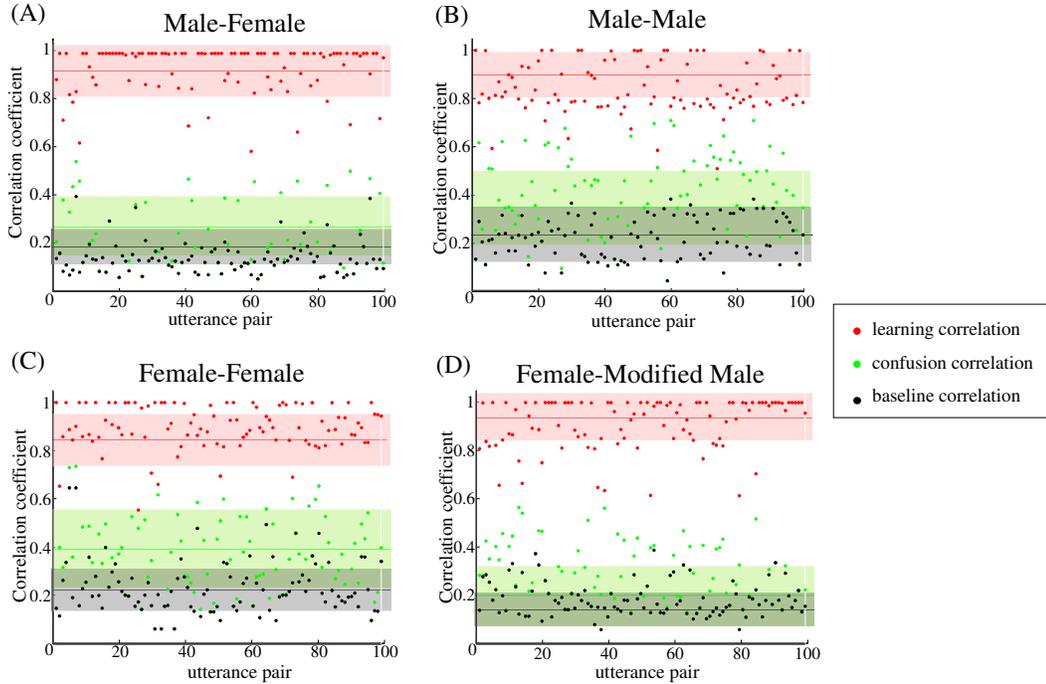


Figure 5.19: Segregating speech utterances using original sentences. **(A)** The statistics for male-female pairs are:  $\bar{\rho}_l = 0.92 \pm 0.1$ ,  $\bar{\rho}_c = 0.27 \pm 0.12$ ,  $\bar{\rho}_b = 0.19 \pm 0.07$ . **(B)** The statistics for male-male pairs are:  $\bar{\rho}_l = 0.9 \pm 0.09$ ,  $\bar{\rho}_c = 0.35 \pm 0.05$ ,  $\bar{\rho}_b = 0.24 \pm 0.11$ . **(C)** The statistics for female-female pairs are:  $\bar{\rho}_l = 0.85 \pm 0.11$ ,  $\bar{\rho}_c = 0.39 \pm 0.16$ ,  $\bar{\rho}_b = 0.2 \pm 0.09$ . **(D)** The statistics for female-modified male pairs are:  $\bar{\rho}_l = 0.94 \pm 0.1$ ,  $\bar{\rho}_c = 0.2 \pm 0.12$ ,  $\bar{\rho}_b = 0.14 \pm 0.07$ .

sentence and the competing sentence. We call this correlation  $\rho_c$ , or confusion correlation. The values obtained for each speaker pair are shown in Figure 5.19, with the mean value shown with a straight line, and the variance shown with a shaded box. The results lead to quite a remarkable correspondence between the original and learned sentences, with correlation coefficient as high as 0.94. We re-display these values in Figure 5.20, but only plotting the results from the 50 best pairs, i.e. the pairs for which we achieved the best separation. This figure confirms how successful the model’s performance is with mean correlation values between 0.95–1.

A similar set of tests was also performed to test segregation of speech versus music

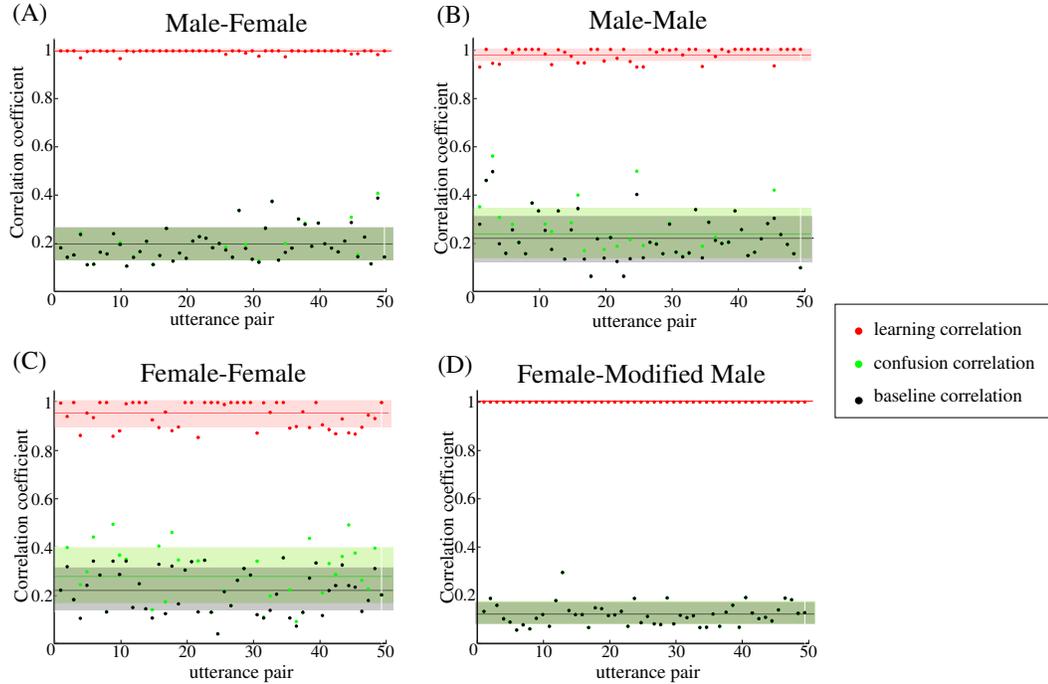


Figure 5.20: Segregating speech utterances using original sentences: 50 best pairs from Figure 5.19. **(A)** The statistics for male-female pairs are:  $\bar{\rho}_l = 1 \pm 0.01$ ,  $\bar{\rho}_c = 0.19 \pm 0.07$ ,  $\bar{\rho}_b = 0.19 \pm 0.07$ . **(B)** The statistics for male-male pairs are:  $\bar{\rho}_l = 0.98 \pm 0.03$ ,  $\bar{\rho}_c = 0.24 \pm 0.10$ ,  $\bar{\rho}_b = 0.22 \pm 0.09$ . **(C)** The statistics for female-female pairs are:  $\bar{\rho}_l = 0.95 \pm 0.05$ ,  $\bar{\rho}_c = 0.28 \pm 0.11$ ,  $\bar{\rho}_b = 0.22 \pm 0.09$ . **(D)** The statistics for female-modified male pairs are:  $\bar{\rho}_l = 1$ ,  $\bar{\rho}_c = 0.12 \pm 0.05$ ,  $\bar{\rho}_b = 0.12 \pm 0.05$ .

melodies. The model was successful in separating these two patterns into separate clusters, particularly when the musical melody consisted of a tune from a single instrument. Musical pieces from an entire orchestra would require a more elaborate restructuring of the model under more complex perceptual principles enabling the identification of musical sounds from many instruments as one perceptual stream, rather than a collection of individual instrumental streams.

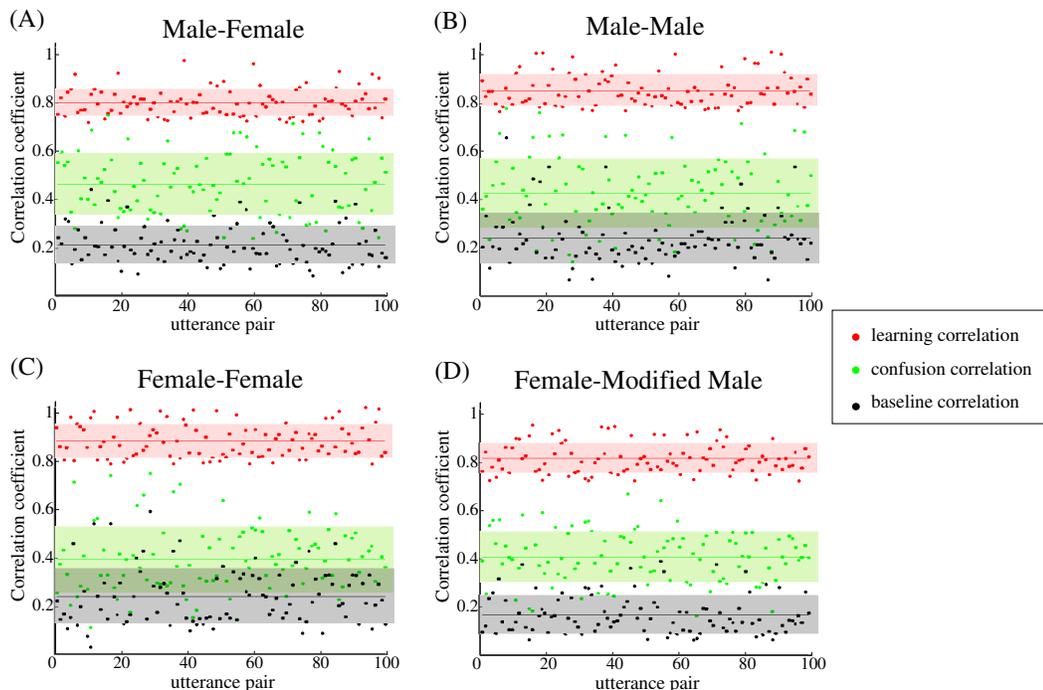


Figure 5.21: Segregating speech mixtures. **(A)** The statistics for male-female pairs are:  $\bar{\rho}_l = 0.78 \pm 0.06$ ,  $\bar{\rho}_c = 0.45 \pm 0.12$ ,  $\bar{\rho}_b = 0.21 \pm 0.08$ . **(B)** The statistics for male-male pairs are:  $\bar{\rho}_l = 0.84 \pm 0.06$ ,  $\bar{\rho}_c = 0.42 \pm 0.14$ ,  $\bar{\rho}_b = 0.24 \pm 0.1$ . **(C)** The statistics for female-female pairs are:  $\bar{\rho}_l = 0.87 \pm 0.07$ ,  $\bar{\rho}_c = 0.83 \pm 0.13$ ,  $\bar{\rho}_b = 0.24 \pm 0.11$ . **(D)** The statistics for female-modified male pairs are:  $\bar{\rho}_l = 0.8 \pm 0.06$ ,  $\bar{\rho}_c = 0.4 \pm 0.1$ ,  $\bar{\rho}_b = 0.16 \pm 0.08$ .

### Speech segregation using utterance mixtures

In a second set of simulations, we employ the adaptive model in its entirety, by involving the pitch and onset extraction modules. In this case, the input signal consists of a sound mixture obtained by summing a pair of sentences from two different speakers. The analysis is then performed on the sound mixture itself, by first extracting pitch and onset elements. Depending on the saliency of each of these vectors, we also keep the original mixture spectrum as additional vector for training the cortical model. In the absence of any pitch or onset evidence, the original spectrum is then kept to represent any additional evidence about the sound patterns at that time instant. For instance, a fricative (such as [s]) would produce an aperiodic hissing that would not produce any salient pitch estimate.

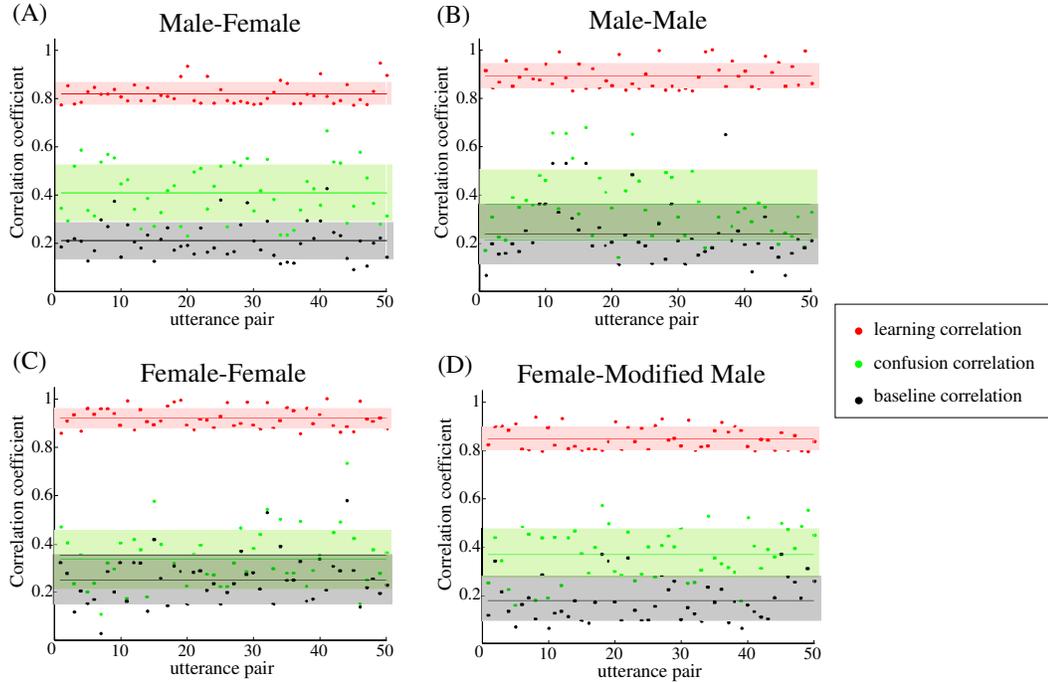


Figure 5.22: Segregating speech mixtures: 50 best pairs from Figure 5.21. **(A)** The statistics for male-female pairs are:  $\bar{\rho}_l = 0.82 \pm 0.04$ ,  $\bar{\rho}_c = 0.41 \pm 0.12$ ,  $\bar{\rho}_b = 0.21 \pm 0.08$ . **(B)** The statistics for male-male pairs are:  $\bar{\rho}_l = 0.89 \pm 0.05$ ,  $\bar{\rho}_c = 0.36 \pm 0.14$ ,  $\bar{\rho}_b = 0.24 \pm 0.12$ . **(C)** The statistics for female-female pairs are:  $\bar{\rho}_l = 0.92 \pm 0.04$ ,  $\bar{\rho}_c = 0.33 \pm 0.12$ ,  $\bar{\rho}_b = 0.25 \pm 0.1$ . **(D)** The statistics for female-modified male pairs are:  $\bar{\rho}_l = 0.85 \pm 0.05$ ,  $\bar{\rho}_c = 0.37 \pm 0.11$ ,  $\bar{\rho}_b = 0.18 \pm 0.09$ .

The outcome of these simulations is given in Figure 5.21. Figure 5.22 again shows the best clustered 50 pairs from the original data set in Figure 5.21. Overall, while the outcome of such simulations leads to more reduced correlation values than those obtained in the earlier case with the original utterances, we are still able to achieve a correspondence of 0.8–0.9, which is rather remarkable using quite little evidence about the signal, consisting of only pitch, onset and original patterns.

## 5.6 Summary and Discussion

In this chapter, we presented a computational approach to the problem of auditory scene analysis inspired from computational strategies of cortical sound processing. Based on the notion of an internal world representation maintained by the cortex, we formulated a sound organization scheme regulated by the statistical principles of Kalman filtering and neural network rules of competitive learning. The model builds and maintains an internal representation of sound streams available in the environment. This representation serves both as a reference set for clustering sensory inputs into their corresponding perceptual streams, as well as an adjustable judgement measure of the auditory scene, whose values are adapted to reflect the variability in the sound streams. The model is founded on perceptual principles of auditory grouping and stream formation. Such principles are translated into a computational scheme that combines aspects of bottom-up primitive sound processing with a top-down internal representation of the world, which adapts its intrinsic representation based on the residual error between its own predictions and the actual sensory input.

The model was tested under a variety of auditory streaming conditions, leading to a successful outcome of these simulations in accord with the responses expected from listening tests. The system could account for the relationship between sequential streaming and various grouping factors, such as frequency separation, presentation rate, timbre differences, background on foreground effects. It also proved successful in reflecting the role of simultaneous processes in separating concurrent sound patterns into their corresponding perceptual streams. Furthermore, the model achieved a very remarkable performance when tested with natural speech mixtures. It successfully identifies and segregates sound elements corresponding to different speakers based on their spectro-temporal patterns,

reflected in the variability of pitch and vocal tract parameters between speakers, as well as the temporal dynamics of the utterances.

What sets this model apart from existing computational approaches to auditory scene analysis is a combination of components which stem from its biological foundation and its computational scheme:

- **Data-driven cortical representation:** The model relies on an initial stage of data processing, that extracts “primitive” acoustic features from the sound mixture. Along with the common cues generally invoked in most CASA systems (e.g., frequency proximity, onset synchrony, harmonicity), the model also explores more high-level data mappings that expose differences between sound elements in a multi-dimensional representation. Physiological evidence of neural response patterns in central auditory areas shows the emergence of selectivity to particular spectral features at the level of pre-cortical and cortical neurons. Such selectivity is much more complex than simple tuning-curves in auditory nerve fibers. We exploit such organization in implementing a wavelet-based multi-scale representation of sound spectra, which successfully captures the variability in sound elements beyond a simple time-frequency mapping. The model is hence given an edge when dealing with cases of interleaved complex pattern sequences (such as alternating ripples), as well as segregation of speaker and sound characteristics in a way to enhance the differences between their voice features (pitch, length of vocal tract, timbre, etc...).
- **Cortical dynamics:** Exploring the role of cortical dynamics in auditory scene analysis is invoked in the model by using a bank of tuned rate-filters within each cluster. Rate selectivity ensures a distribution of sound elements within each cluster depending on their temporal dynamics, and guarantees a fidelity of representation

of changes along the time dimension in the response patterns of the filters. For instance, in the streaming simulation of alternating A-B-A-B... tones (Figure 5.8), the presence of a 4Hz selective filter whose time constant is commensurate with the presentation rate of the sequence ensures both responsiveness of the model to that specific presentation speed, in addition to maintaining a non-interrupted response pattern reflecting the perceived *continuous* A and B sounds once streaming effects take place. Furthermore, the temporal dynamics within each cluster act as a memory component to reflect past or already-learned features of the different sound streams, as well as projecting them into the future.

- **Kalman filtering:** Kalman estimation sets a rigorous framework for robust inference of implicit information about the state of the system, reflecting the sensory input from the environment. The simplicity and recursive nature of Kalman filtering are some of its appealing properties – it makes real-time implementations of the system much more feasible, and sets an attractive approach to computationally implement *sequential* processing phenomena [21] in an optimal and straightforward fashion . Along with its well-established benefits, Kalman filter’s real contribution to the current model is its ability to track the changes of the filters’ states in time. As the sound elements in the acoustic mixture vary with time, and given the constraints on the model to be able to learn the sound structures “on-the-fly”, it is of paramount importance that the model robustly tracks these temporal changes. Such requirement is mostly valuable in cases of speech and non-stationary signals where temporal progression in time is an important clue to the organization of sounds into separate streams.

- **Interaction between primitive and top-down influences:** The model offers a natural and biologically plausible framework for interactions between primitive and internal representations of acoustic elements. For cortical structures to be able to maintain accurate estimates of auditory events, they have to reconcile between their current states and the input from the environment. This interaction can be understood in the context of adaptation of cortical structures to the requirements of the external world along with predictive anticipation or cognitive influence of the context in which the sounds are being segregated or fused. For instance, the choice of the limiting boundaries between the different clusters can be reflected in the choice of a distance measure (chosen to be euclidian distance in the current model, Equation 5.6), which is an element that can be further explored in enforcing specific “schema-based” processes into the “primitive” segregation of sounds.

Along with the main contributions of this model, we can reflect on the challenges faced in the implementation and application of such system. As far as failures of the model are concerned, they tend to occur mostly in cases where ambiguous and complex sound patterns are being analyzed. For instance, specific situations of orchestral musical scenes raise questions as to the level of granularity at which the system should operate to separate the instrumental and background elements from the scene. Such decision does not have an obvious answer, even theoretically, since our individual expectations and familiarity with the musical melody and instruments define how we would perceive such a complex scene. In the absence of more specific system modules reflecting prior experiences or learned knowledge, one does not expect a general-purpose auditory organization system such as the one presented in this chapter to be successful in handling such complex ambiguous cases.

The model is also faced with a challenging task of “learning” through a single pass of the data. The system is in fact expected to “get it right” in one trial, with no repetition or exposure to information from a given time instant more than once. The one-time requirement is necessary to maintain the flow of temporal continuity constraints on the learning process. Presenting the same time instants twice would disrupt such continuity, and hence defeats the purpose of enforcing smoothness constraints on the learning function. Nonetheless, the system proves to be quite successful despite this challenging aspect. It also demonstrates that such model can be highly invaluable for real-time applications, particularly interfacing with automatic speech recognition systems, hearing prostheses, as well as general sound separation and enhancement applications. One of the obvious shortcomings of the model however is the difficulty to re-synthesize the perceived streams given the current structure of the system. Simplification of the computational load is probably the best approach in alleviating the intense computational complexity of the model. Nonetheless, the model is extremely valuable in exploring various aspects of sound organization in the brain, allowing us to gain interesting insight into the neural basis of auditory scene analysis.

The current model has interesting implications in suggesting a cortical basis for sound organization and scene analysis. The auditory cortex has been previously probed for evidence of its role in auditory streaming and sound segregation. Mechanisms such as synaptic depression [4] and adaptation or forward masking [14, 68] have been presented as plausible neural mechanisms for explaining the perceptual effects of auditory streaming. In the current model, we speculate that the very structure of response patterns in the cortex, along with its tuned selectivity to specific sound features is also important in understanding the cortical role of auditory stream formation and sound segregation. These

properties are not necessarily contradictory with the explanations just mentioned about a role of synaptic properties in cortical auditory streaming. We have indeed shown in Chapter 3 that such neural mechanisms can in fact be attributed a role in explaining the computational characteristics of cortical processing. The emergence of its unique temporal properties is in fact in agreement with the known perceptual attributes of auditory scene analysis.

## Chapter 6

# Conclusion

### 6.1 Thesis overview

As we described the intricacy of the hearing problem and its multiple facets, our sense of the complexity of tackling sound processing problems is only reinforced. In this thesis, we focused on two directions of research pertaining to sound processing: neural and theoretical modelling. The goal of the current work is to investigate the neural mechanisms underlying auditory perception of complex sounds at the cortical level, and to formalize computational models of these mechanisms. Such models can then be the backbone for simulating the cognitive function of brain in artificial systems.

Our interest in focusing on the neural basis of auditory processing is not to formalize a canonical map of the cortex. We are instead more interested in learning the computational strategies that govern the cortical function. In Chapter 3, we focused on a

particularly important component of neural processing in audition, that of temporal encoding of sound. Evidence shown in this work argues for an interesting interplay between multiple time scales that determine neural responses in the primary auditory cortex (A1). While evidence for differential temporal encoding of sound has been known for decades, it has often been assumed that cortical neurons are mainly characterized by their slow response dynamics focusing on overall spectrotemporal patterns in sound that are most relevant for speech and music perception. Here, we presented important physiological evidence of cortical capability of showing precise time locking to sound features as low as 1msec. Such interplay between cortical time constants hints to a role of A1 in integration of sound features at different scales, and possibly in the formation of auditory objects.

By reviewing various ingredients of physiological sound processing, we were able to explore computational implementations of these principles in tackling specific perceptual questions. In Chapter 4, we described a computational approach to evaluating the question of speech intelligibility. We demonstrated that mapping sound features into a spectrotemporal modulation space can accurately predict the intelligibility level of a signal under various noise conditions. This algorithmic implementation of this concept defines a new intelligibility metric (called the STMI, Spectro-Temporal Modulation Index) which was validated by performing a psychoacoustic study to correlate human estimates with model predictions.

We also exploited the neural evidence from studying cortical responses in formalizing a general scheme for auditory scene analysis, where we abstract perceptual and physiological principles of auditory processing into a statistical machine learning model. The approach, described in Chapter 5 outlines an adaptive learning model of sound organization into separate perceptual streams, based on a cortical model of dynamic filtering. It

is based on reconciling evidence between observed sensory inputs and predictions of an internal world representation, which reflects the dynamics of the cortical learning module. The algorithm presents no assumption as far as the sound elements are concerned, and relies on a dual operation between Kalman-based estimation and unsupervised learning to organize sound elements into their corresponding perceptual streams. The model is shown to be quite successful in reproducing streaming effects previously tests with human subjects, as well as addressing the more practical question of speaker separation.

In addition to their scientific contribution, the models described in Chapters 4 and 5 find a direct applicability in many engineering problems, such as hearing-aid research (where current collaboration work is under way for relating *STMI* estimates with behavioral advantages of particular microphones in hearing-aid circuits), room acoustics and communication channels, automatic sound processing, etc.

As our scientific curiosity drives us to explore the whole of “audition”, we find ourselves facing a tremendously complex and intertwined problem that only collaborative work can hope to address without falling into naive simplifications and oversights of the bigger questions. While a considerable amount of effort is being put forward in the scientific community, especially as far as modelling work is concerned, most of the studies remain independent of each other, setting their own rules and simplifications of the perceptual theory. Such approach hinders the progress towards generalized theories of auditory perception, and hence the development of practical engineering tools for sound processing. As our physiological knowledge expands with further investigation from experimental neuroscience, the need for better interpretations of the computational task of biological sound processing becomes even greater. The gap between theoretical and experimental neuroscience can only be bridged through improved and well-founded theories of

the hearing problem, allowing the models to anticipate predictions that can be tested by biological experiments, and setting the perfect interaction in collaborative efforts between theory, experimentation, and engineering application.

## 6.2 Future prospects

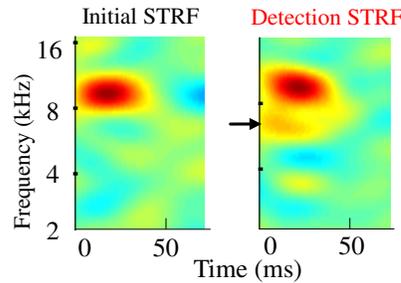


Figure 6.1: Receptive field patterns of a cortical neuron during a *passive* state are modified when the animal is engaged in a task that requires to detect the presence of a tone frequency at 6KHz (shown by the black arrow in the right panel). The neural selectivity is potentiated to enhance the responsiveness of this neuron to a 6KHz frequency (STRF in right panel), hence increasing the likelihood of capturing the attended 6KHz target during the acoustic task (from [70]).

By formalizing a model of auditory scene analysis or auditory perception, we see an interesting extension of this work in light of recent data on neural plasticity in the auditory cortex [70]. It is becoming increasingly clear from physiological studies that receptive fields in the auditory cortex are constantly adapting and re-organizing dynamically to meet the challenges of an ever-changing environment and new behavioral demands. These changes can in fact be examined in the context of the interaction between sound inputs and the internal states of the individual neurons, which drive it to anticipate certain patterns from the environment. Directed attention and sound anticipation leads to changes in receptive field properties of individual neurons and the cortical ensemble, which itself plays a role in enhancing the behavioral performance during acoustic tasks (Figure 6.1).

## Appendix A

# Derivation of predictive learning

### A.1 Optimizing the learning function

We start with a posterior optimization function (Equation 5.3), which equivalently corresponds to (Equation 5.4):

$$\begin{aligned}\mathcal{J} &= \max P(\vec{\mathcal{Z}}|\vec{\mathbf{I}}) \\ &= \min_{\alpha} \sum_{\alpha} \left[ (\mathbf{I}^{\beta} - \mathbf{A}\mathcal{Z}^{\alpha})^T \Sigma^{-1} (\mathbf{I}^{\beta} - \mathbf{A}\mathcal{Z}^{\alpha}) + (\vec{\mathcal{Z}}^{\alpha} - \mathcal{Z}^{\alpha})^T \Pi^{-1} (\vec{\mathcal{Z}}^{\alpha} - \mathcal{Z}^{\alpha}) \right]\end{aligned}\tag{A.1}$$

Maximizing  $\mathcal{J}$  corresponds to finding the optimal vector  $\hat{\mathcal{Z}}^{\alpha}$  such that  $\frac{\partial \mathcal{J}}{\partial \mathcal{Z}^{\alpha}} = 0$ .

$$\begin{aligned}
& \frac{\partial \mathcal{J}}{\partial \hat{\mathbf{Z}}^\alpha} = 0 \\
\Rightarrow & -\mathbf{A}^T \Sigma^{-1} (\mathbf{I}^\beta - \mathbf{A} \hat{\mathbf{Z}}^\alpha) + \Pi^{-1} (\hat{\mathbf{Z}}^\alpha - \bar{\mathbf{Z}}^\alpha) = 0 \\
\Rightarrow & (\mathbf{A}^T \Sigma^{-1} \mathbf{A} + \Pi^{-1}) \hat{\mathbf{Z}}^\alpha = \Pi^{-1} \bar{\mathbf{Z}}^\alpha + \mathbf{A}^T \Sigma^{-1} \mathbf{I}^\beta \\
\Rightarrow & \hat{\mathbf{Z}}^\alpha = (\mathbf{A}^T \Sigma^{-1} \mathbf{A} + \Pi^{-1})^{-1} (\Pi^{-1} \bar{\mathbf{Z}}^\alpha + \mathbf{A}^T \Sigma^{-1} \mathbf{I}^\beta) \\
\Rightarrow & \hat{\mathbf{Z}}^\alpha = \bar{\mathbf{Z}}^\alpha + (\mathbf{A}^T \Sigma^{-1} \mathbf{A} + \Pi^{-1})^{-1} \mathbf{A}^T \Sigma^{-1} (\mathbf{I}^\beta - \mathbf{A} \bar{\mathbf{Z}}^\alpha) \\
\Rightarrow & \hat{\mathbf{Z}}^\alpha = \bar{\mathbf{Z}}^\alpha + \mathbf{G} (\mathbf{I}^\beta - \mathbf{A} \bar{\mathbf{Z}}^\alpha)
\end{aligned} \tag{A.2}$$

where  $\mathbf{G} \triangleq (\mathbf{A}^T \Sigma^{-1} \mathbf{A} + \Pi^{-1})^{-1} \mathbf{A}^T \Sigma^{-1}$ . The matrix inversion of  $(\mathbf{A}^T \Sigma^{-1} \mathbf{A} + \Pi^{-1})$  is performed under assumptions of non-singularity. Such assumption can be guaranteed by a proper choice of the state matrix  $\mathbf{A}$  and the noise covariance matrices  $\Sigma$  and  $\Pi$ .

## A.2 Difference to state-space equation

In this section, we outline the steps involved in converting a difference equation into a state-space equation. Consider the general difference equation:

$$a_0 Y(t) + a_1 Y(t-1) + \cdots + a_n Y(t-n) = b_0 X(t) + b_1 X(t-1) + \cdots + b_{n-1} X(t-n+1) \tag{A.3}$$

The conversion into a state-space model can be achieved under two schemes: (1) a controllable canonical state-space representation, and (2) an observable canonical form. The difference between the two representation varies simply in the definition of the state or internal variables, and can be chosen differently depending on the application [123]. In the current work, we follow the derivation method based on the controllable state-space

representation, as it gives a more intuitive interpretation of what the internal variables are. We define the state variable to correspond to the state of the delay registers in a direct form II representation of the difference equation [115]. We define the state variables as:

$$\begin{aligned}\mathcal{Z}(t) &\triangleq Y(t) - \{b_1X(t-1) + \dots + b_nX(t-n)\} \\ X(t) &= a_0\mathcal{Z}(t) + a_1\mathcal{Z}(t-1) + \dots + a_{n-1}\mathcal{Z}(t-n+1)\end{aligned}\tag{A.4}$$

From this representation, we define a vector of state variables

$\vec{\mathcal{Z}}(t) = [\mathcal{Z}(t)\mathcal{Z}(t-1)\dots\mathcal{Z}(t-n+1)]^T$ , where  $T$  is the transpose operator. By rewriting Equation A.4 in terms of delayed versions of the state variables  $\mathcal{Z}$ , we get:

$$\mathcal{Z}(t) = Y(t) - \{b_1\mathcal{Z}(t-1) + \dots + b_n\mathcal{Z}(t-n)\}\tag{A.5}$$

which can also be rewritten in vector form, as:

$$\begin{bmatrix} \mathcal{Z}(t) \\ \mathcal{Z}(t-1) \\ \vdots \\ \mathcal{Z}(t-n+1) \end{bmatrix} = \begin{bmatrix} -b_1 & \dots & -b_{n-1} & -b_n \\ 1 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{Z}(t) \\ \mathcal{Z}(t-1) \\ \vdots \\ \mathcal{Z}(t-n+1) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} Y(t)\tag{A.6}$$

and  $X(t)$  is defined by:

$$X(t) = a_0\mathcal{Z}(t) + \dots + a_{n-1}\mathcal{Z}(t-n+1)\tag{A.7}$$

The final state-space equations are give by:

$$\vec{Z}(t) = \mathbf{B}\vec{Z}(t-1) + \mathbf{C}Y(t) \quad (\text{A.8})$$

$$X(t) = \mathbf{A}\vec{Z}(t) \quad (\text{A.9})$$

where,

$$\mathbf{A} \triangleq [a_0 \cdots a_{n-1}] \quad \text{and} \quad \mathbf{B} \triangleq \begin{bmatrix} -b_1 & \cdots & -b_{n-1} & -b_n \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{C} \triangleq \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

## BIBLIOGRAPHY

- [1] The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom. MIT, SRI International, Texas Instruments.
- [2] International electrotechnical commission. <http://www.iec.ch/>.
- [3] International organization for standardization. <http://www.iso.org/>.
- [4] L. F. Abbott, K. Sen, J. A. Varela, and S. B. Nelson. Synaptic depression and cortical gain control. *Science*, 275:220–224, 1997.
- [5] M. Abeles. *Local cortical circuits: an electrophysiological study*. Springer, Berlin, 1981.
- [6] American National Standards Institute, New York. *American national standard methods for calculation of the speech intelligibility index*, 1997. ANSI S3.5.
- [7] B. O. Anderson and J. B. Moore. *Optimal filtering*. Information and system sciences series. Prentice-Hall, Inc., New Jersey, 1979.

- [8] W. Bair and C. Koch. Temporal precision of spike trains in extrastriate cortex of behaving macaque monkey. *Neural computation*, 8:1185–1202, 1996.
- [9] O. Bar-Yosef, Y. Rotman, and I. Nelken. Responses of neurons in cat primary auditory cortex to bird chirps: effects of temporal and spectral context. *Journal of Neuroscience*, 22:8619–8632, 2002.
- [10] H. Barlow. *Large-Scale Neuronal Theories of the Brain*, chapter What is the computational goal of neocortex, pages 1–22. MIT Press, Cambridge, MA, 1994.
- [11] M. F. Bear, B. W. Connors, and M. A. Paradiso. *Neuroscience: exploring the brain*. Williams and Wilkins, 1996.
- [12] M. W. Beauvois and R. Meddis. A computer model of auditory stream segregation. *The Quarterly Journal of Experimental Psychology*, 43A(3):517–541, 1991.
- [13] S. Becker and M. Plumbley. Unsupervised neural network learning procedures for feature extraction and classification. *International Journal of Applied Intelligence*, 6(3):185–203, 1996.
- [14] M. A. Bee and G. M. Klump. Primitive auditory stream segregation: A neurophysiological study in the songbird forebrain. *Journal of Neurophysiology*, 92:1088–1104, 2004.
- [15] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind source separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.
- [16] J. C. Bellamy. *Digital telephony*. Wiley series in telecommunications and Signal Processing. John Wiley and Sons, New York, third edition, 2000.

- [17] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 1981.
- [18] A. Bieser and P. Müller-Preuss. Auditory responsive cortex in the squirrel monkey: neural responses to amplitude-modulated sounds. *Experimental Brain Research*, 108:273–284, 1996.
- [19] S. A. Billings and W. S. F. Voon. Piecewise linear identification of nonlinear systems. *International journal of control*, 46(1):215–235, 1987.
- [20] J. S. Bradley. Predictors of speech intelligibility in rooms. *The Journal of the Acoustical Society of America*, 80(3):837–845, 1986.
- [21] A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT Press, 1990.
- [22] A. S. Bregman and P. A. Ahad. Demonstrations of auditory scene analysis: The perceptual organization of sound. Compact Disk, Department of Psychology, McGill University.
- [23] G. J. Brown. *Computational auditory scene analysis: A representational approach*. PhD thesis, University of sheffield, 1992.
- [24] G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8(4):297–336, 1994.
- [25] G. J. Brown and M. P. Cooke. *Computational auditory scene analysis*, chapter Temporal synchronization in a neural oscillator model of primitive auditory stream segregation. Lawrence erlbaum associates, London, 1998.

- [26] M. Carandini, D. J. Heeger, and W. Senn. A synaptic explanation of suppression in visual cortex. *Journal of Neuroscience*, 22:10053–10065, 2002.
- [27] R. P. Carlyon, R. Cusack, J. M. Foxtan, and I. H. Robertson. Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 27:115–127, 2001.
- [28] R. P. Carlyon, J. Deeks, D. Norris, and S. Butterfield. The continuity illusion and vowel identification. *Acta Acustica - Acustica*, 88:408–415, 2002.
- [29] F. S. Chance, S. B. Nelson, and L. F. Abbott. Synaptic depression and the temporal response characteristics of v1 cells. *Journal of Neuroscience*, 18:4785–4799, 1998.
- [30] J. A. Cherry. Distortion analysis of weakly nonlinear filters using volterra series.
- [31] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. A. Shamma. Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America*, 106(5):2719–2732, 1999.
- [32] T. Chi, P. Ru, and S. A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *Speech Communication*, 2003.
- [33] C. K. Chui and G. Chen. *Kalman filtering with real time applications*. Springer-Verlag, third edition, 1999.
- [34] S. Chung, X. Li, and S. B. Nelson. Short-term depression at thalamocortical synapses contributes to rapid adaptation of cortical sensory responses in vivo. *Neuron*, 34:437–446, 2002.
- [35] P. Churchland, V. S. Ramachandran, and T. Sejnowski. *Large-scale neuronal theories of the brain*, chapter A critique of pure vision. MIT Press, 1994.

- [36] M. A. Cohen, S. Grossberg, and L. L. Wyse. A spectral network model of pitch perception. *Journal of the Acoustical Society of America*, 98:862–879, 1995.
- [37] M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35:141–177, 2001.
- [38] M. P. Cooke. *Modeling auditory processing and organisation*. PhD thesis, University of sheffield, 1991.
- [39] O. Creutzfeldt, F. C. Hellweg, and C. Schreiner. Thalamocortical transformation of responses to complex auditory stimuli. *Experimental Brain Research*, 39:7–104, 1980.
- [40] G. L. Dannenbring and A. S. Bregman. Stream segregation and the illusion of overlap. *Journal of Experimental Psychology: Human Perception and Performance*, 2:544–555, 1976.
- [41] C. J. Darwin. Auditory grouping and attention to speech. In *proceedings of the Institute of Acoustics*, volume 23, pages 165–172, 2001.
- [42] C. J. Darwin and R. P. Carlyon. *The handbook of perception and cognition*, volume 6, chapter Auditory grouping, pages 387–424. Academic Press, 1995.
- [43] T. Dau, D. Püschel, and D. Kohlrausch. A quantitative model of the effective signal processing in the auditory system. *The Journal of the Acoustical Society of America*, 99(6):3615–3622, 1996.
- [44] A. de Cheveigné. Time-domain auditory processing of speech. *Journal of Phonetics*, 31:547–561, 2003.

- [45] E. DeBoer. *Handbook of sensory physiology*, chapter On the residue in hearing and auditory pitch perception, pages 479–583. Berlin: Springer, 1976.
- [46] E. DeBoer. *Time Resolution in Auditory Systems*, chapter Auditory time constants: A paradox?, pages 141–158. Springer-Verlag, Berlin, 1985.
- [47] R. C. DeCharms, D. T. Blake, and M. M. Merzenich. Optimizing sound features for cortical neurons. *Science*, 280:1439–1443, 1998.
- [48] S. L. Denham. *Computational models of auditory function*, volume 312, chapter Cortical synaptic depression and auditory perception, pages 281–296. NATO Science Series, Life Science, IOS, Amsterdam, 2001.
- [49] 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. *Journal of Neurophysiology*, 85:1220–1234, 2001.
- [50] F. deRibaupierre, M. H. Goldstein, and G. Yeni-Komshian. Cortical coding of repetitive acoustic pulse. *Brain Research*, 48:205–225, 1972.
- [51] F. deRibaupierre, M. H. Goldstein, and G. Yeni-Komshian. Intracellular study of the cat’s primary auditory cortex. *Brain Research*, 48:185–204, 1972.
- [52] F. deRibaupierre and E. Rouiller. Temporal coding of repetitive clicks : Presence of rate selective units in cat’s medial geniculate body (MGB). *Journal of Physiology*, 318:23–24, 1981.
- [53] R. Drullman, J. Festen, and R. Plomp. Effect of envelope smearing on speech perception. *The Journal of the Acoustical Society of America*, 95(2):1053–1064, 1994.

- [54] J. D. Durrant and J. H. Lovrinic. *Bases of hearing science*. Williams and Wilkins, baltimore, third edition, 1995.
- [55] J. J. Eggermont. Between sound and percpetion: reviewing the search for a neural code. *Hearing Research*, 157:1–42, 2001.
- [56] J. J. Eggermont. Temporal modulation transfer functions in cat primary auditory cortex: Separating stimulus effects from neural mechanisms. *Journal of Neurophysiology*, 87:305–321, 2002.
- [57] J. J. Eggermont, P. M. Johannesma, and A. M. Aertsen. Reverse correlation methods in auditory research. *Quarterly review of Biophysics*, 16(3):341–414, 1983.
- [58] J. J. Eggermont and G. Smith. Characterizing auditory neurons using the Wigner and Rihacek distributions: A comparison. *Journal of Acoustical Society of America*, 87:246–259, 1990.
- [59] M. Elhilali, T. Chi, and S. A. Shamma. A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication*, 41:331–348, 2003.
- [60] M. Elhilali, J. B. Fritz, D. Bozak, D. A. Depireux, and S. A. Shamma. Comparison of response characteristics in auditory cortex of the awake and anesthetized ferret. *Association of Research in Otolaryngology Abstracts*, 25(162), 2002.
- [61] M. Elhilali, J. B. Fritz, D. J. Klein, J. Z. Simon, and S. A. Shamma. Dynamics of precise spike timing in primary auditory cortex. *Journal of Neuroscience*, 24(5):1159–1172, 2004.
- [62] M. Elhilali, D. J. Klein, J. B. Fritz, J. Z. Simon, and S. A. Shamma. *The enigma of cortical responses: slow yet precise*, pages 403–409. Springer, New York, 2005.

- [63] D. P. W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996.
- [64] A. K. Engel and W. Singer. Neuronale grundlagen der gestaltwahrnehmung. In *Spektrum der Wissenschaft, Dossier 4/97, "Kopf und Computer"*, pages 66–73. Spektrum Akademischer Verlag, Heidelberg, 1997.
- [65] M. A. Escabí and C. E. Schreiner. Non-linear spectro-temporal envelope processing in cat ICC. *Association of Research in Otolaryngology Abstracts*, 22(869), 1999.
- [66] A. F. Fahn and J. Santos-Sacchi, editors. *Physiology of the ear*. Singular, thomson learning, second edition, 2001.
- [67] D. Ferster. Linearity of synaptic interactions in the assembly of receptive fields in cat visual cortex. *Current Opinion in Neurobiology*, 4:563–568, 1994.
- [68] Y. I. Fishman, D. H. Reser, J. C. Arezzo, and M. Steinschneider. Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hearing Research*, 151:167–187, 2001.
- [69] J. B. Fritz, D. A. Depireux, J. Z. Simon, and S. A. Shamma. Neuronal response characteristics in auditory cortex of the awake ferret. *Association of Research in Otolaryngology Abstracts*, 24, 2001.
- [70] J. B. Fritz, S. A. Shamma, M. Elhilali, and D. J. Klein. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, 6(11):1216–1223, 2003.
- [71] Z. Gil, Y. Amitai, M. A. Castro, and B. W. Connors. Differential regulation of neocortical synapses by neuromodulators and activity. *Neuron*, 19:679–686, 1997.

- [72] H. Gockel, R. P. Carlyon, and J. M. Deeks. Effect of modulator asynchrony of sinusoidal and noise modulators on frequency and amplitude modulation detection interference. *Journal of the Acoustical Society of America*, 112(6):2975–2984, 2002.
- [73] J. L. Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54:1496–1516, 1973.
- [74] S. Greenberg and T. Arai. Speech intelligibility is highly tolerant for cross-channel spectral asynchrony. In *Proceedings of the Joint Meeting of the Acoustical Society of America and the International Congress on Acoustics, Seattle*, pages 2677–2678, 1998.
- [75] N. Grimault, S. P. Bacon, and C. Micheyl. Auditory stream segregation on the basis of amplitude-modulation rate. *Journal of the Acoustical Society of America*, 111:1340–1348, 2002.
- [76] B. Blankertz H. Purwins and K. Obermayer. Computing auditory perception. *Organised Sound*, 5(3):159–171, 2000.
- [77] W. M. Hartman and D. Jonhson. Stream segregation and peripheral channelling. *Music Perception*, 9(2):155–184, 1991.
- [78] M. J. Hawken and D. H. Grosop R. M. Shapley. Temporal-frequency selectivity in monkey visual cortex. *Visual Neuroscience*, 13:477–492, 1996.
- [79] P. Heil. Auditory cortical onset responses revisited. i. first-spike timing. *Journal of Neurophysiology*, 77:2616–2641, 1997.

- [80] D. Hermes, A. Aertsen, P. Johannesma, and J. Eggermont. Spectro-temporal characteristics of single units in the midbrain of the lightly anaesthetised grass frog (*Rana temporaria* L.) investigated with noise stimuli. *Hearing Research*, 5:147–178, 1981.
- [81] D. J. hermes and J. C. vanGestel. The frequency scale of speech intonation. *The Journal of the Acoustical Society of America*, 90(1):97–102, 1991.
- [82] J. A. Hirsch, J. M. Alonso, J. C. Reid, and L. M. Martinez. Synaptic integration in striate cortical simple cells. *Journal of neuroscience*, 18:9517–9528, 1998.
- [83] T. Houtgast and H. J. M. Steeneken. Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics. *Acoustica*, 46:60–72, 1980.
- [84] T. Houtgast and H. J. M. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77(3):1069–1077, 1985.
- [85] F. V. Jensen. *Bayesian networks and decision diagrams*. Springer, 2001.
- [86] P. Julià, A. Desages, and O. Agamennoni. High level canonical piecewise linear representation using simplicial partition. *IEEE transactions on circuits and systems - I: Fundamental theory and applications*, 46(4):463–480, 1999.
- [87] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, editors. *Principles of neural science*. McGraw-Hill, New York, fourth edition, 2000.
- [88] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma. Robust spectro-temporal reverse correlation for the auditory system: optimizing stimulus design. *Journal of Computational Neuroscience*, 9:85–111, 2000.

- [89] M. J. Korenberg and I. W. Hunter. The identification of nonlinear biological systems: Volterra kernel approaches. *Annals of Biomedical Engineering*, 24(2):250–268, 1996.
- [90] N. Kowalski, D. A. Depireux, and S. A. Shamma. Analysis of dynamic spectra in ferret primary auditory cortex. i. characteristics of single-unit responses to moving ripple spectra. *Journal of Neurophysiology*, 76(5):3503–3523, 1996.
- [91] N. Kowalski, D. A. Depireux, and S. A. Shamma. Analysis of dynamic spectra in ferret primary auditory cortex. ii. prediction of unit responses to arbitrary dynamic spectra. *Journal of Neurophysiology*, 76(5):3524–3534, 1996.
- [92] A. E. Krukowski and K. D. Miller. Thalamocortical nmda conductances and intracortical inhibition can explain temporal tuning. *Nature Neuroscience*, 4:424–430, 2001.
- [93] K. D. Kryter. Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, 34(11):1689–1697, 1962.
- [94] G. Langner. Periodicity coding in the auditory system. *Hearing Research*, 6:115–142, 1992.
- [95] E. A. Lee and D. G. Messerschmitt. *Digital Communication*. Kluwer Academic Publishers, Boston, second edition, 1994.
- [96] M. S. Lewicki. Bayesian modeling and classification of neural signals. *Neural Computation*, 6:1005–1030, 1994.
- [97] J. C. R. Licklider. Periodicity pitch and place pitch. *Journal of the Acoustical Society of America*, 26:945, 1954.

- [98] L. Ljung. *System identification: Theory for the user*. Prentice Hall Inc., NJ, second edition, 1999.
- [99] T. Lu, L. Liang, and X. Wang. Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nature Neuroscience*, 4:1131–1138, 2001.
- [100] R. Lyon and S. A. Shamma. *Auditory Computation*, volume 6 of *Springer Handbook of Auditory Research*, chapter Auditory representations of timbre and pitch, pages 221–270. Springer-Verlag New York, Inc., 1996.
- [101] C. Machens, M. Wehr, and A. Zador. Linearity of cortical receptive fields measured with natural sounds. *Journal of Neuroscience*, 24:1089–1100, 2004.
- [102] D. Marr. *Vision*. W. H. Freeman and Company, 1982.
- [103] S. L. McCabe and M. J. Denham. A model of auditory streaming. *Journal of the Acoustical Society of America*, 101(3):1611–1621, 1997.
- [104] M. F. McKinney, M. J. Tramo, and B. Delgutte. *Physiological and psychophysical bases of auditory function*, chapter Neural correlates of musical dissonance in the inferior colliculus, pages 83–89. Shaker Publishing: Maastricht, The Netherlands, 2001.
- [105] K. M. McLeod and C. E. Carr. *Synaptic dynamics and intensity coding in the cochlear nucleus*, pages 416–422. Springer, New York, 2005.
- [106] K. D. Miller. Understanding layer 4 of the cortical circuit: a model based on cat v1. *Cerebral Cortex*, 13:73–82, 2003.

- [107] L. M. Miller, M. A. Escabi, H. L. Read, and C. E. Schreiner. Functional convergence of response properties in the auditory thalamocortical system. *Neuron*, 32:151–160, 2001.
- [108] L. M. Miller, M. A. Escabi, H. L. Read, and C. E. Schreiner. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal fo Neurophysiology*, 87(1):516–527, 2002.
- [109] B. C. J. Moore, editor. *Hearing*. Handbook of Perception and Cognition. Academic Press, San Diego, second edition, 1995.
- [110] B.C.J. Moore and H. Gockel. Factors influencing sequential stream segregation. *Acta Acustica - Acustica*, 88:320–332, 2002.
- [111] J. A. Movshon, I. D. Thompson, and D. J. Tolhurst. Spatial and temporal contrast sensitivity of neurones in areas 17 and 18 of the cat’s visula cortex. *Journal of Physiology*, 283:101–120, 1978.
- [112] I. Nelken. Processing of complex stimuli and natural scenes in the auditory cortex. *Current Opinion in Neurobiology*, 14:474–480, 2004.
- [113] I. Nelken, A. Fishbach, L. Las, N. Ulanovsky, and D. Farkas. Primary auditory cortex of cats: feature detection or something else? *Biological Cybernetics*, 89:397–406, 2003.
- [114] P. C. Nelson and L. Carney. A physiological model for neural responses to amplitude-modulated stimuli and for psychophysical modulation tuning. *Association of Research in Otalaryngology Abstracts*, 26, 2003.

- [115] A.V. Oppenheim and R.W. Shafer. *Discrete Time Signal Processing*. Prentice Hall Publication, second edition, 1999.
- [116] A. R. Palmer, I. M. Winter, and S. E. Stabler. *Advances in Speech, Hearing and Language Processing*, chapter Responses to simple and complex sounds in the cochlear nucleus of the guinea pig. JAI Press, London, 1995.
- [117] S. Paoletti. *Identification of piecewise affine models*. PhD thesis, University of Siena, 2004.
- [118] A. Papoulis. *The Fourier integral and its applications*. McGraw-Hill book Company, 1962.
- [119] A. Papoulis. *Probability, random variables, and stochastic processes*. McGraw-Hill book Company, 1991.
- [120] T. W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, 66(4):911–918, 1976.
- [121] K. L. Payton and L. D. Braida. A method to determine the speech transmission index from speech waveforms. *The Journal of the Acoustical Society of America*, 106(6):3637–3648, 1999.
- [122] D. J. Pinto, J. A. Hartings, J. C. Brumberg, and D. J. Simons. Cortical damping: Analysis of thalamocortical response transformations in rodent barrel cortex. *Cerebral cortex*, 13(1), 2003.
- [123] D. S. G. Pollock. *A handbook of time-series analysis, signal processing and dynamics*. Academic Press, 1999.

- [124] R. A. Rasch. Perception of simultaneous notes such as in polyphonic music. *Acustica*, 40:1–72, 1978.
- [125] R. E. Remez and P. E. Rubin. On the intonation of sinusoidal sentences: contour and pitch height. *Journal of the Acoustical Society of America*, 94(4):1983–1988, 1993.
- [126] B. Roberts and A. S. Bregman. Effects of the pattern of spectral spacing on the perceptual fusion of harmonics. *Journal of the Acoustical Society of America*, 90(6):3050–3060, 1991.
- [127] D. F. Rosenthal and H. G. Okuno, editors. *Computational auditory scene analysis*. Lawrence erlbaum associates, London, 1998.
- [128] M. Sahani and J. F. Linden. *Advances in neural information processing systems*, chapter How linear are auditory cortical responses?, pages 125–132. Cambridge: MIT, 15 edition, 2003.
- [129] C. E. Schreiner and B. Calhoun. Spectral envelope coding in cat primary auditory cortex: Properties of ripple transfer functions. *Journal of Auditory Neuroscience*, 1:39–61, 1995.
- [130] C. E. Schreiner and J. V. Urbas. Representation of amplitude modulation in the auditory cortex of the cat. ii. comparison between cortical fields. *Hearing Research*, 32:49–64, 1988.
- [131] S. A. Shamma. *Methods of Neuronal modeling*, chapter Lateral inhibition network, pages 411–460. MIT Press, second edition, 1998.

- [132] S. A. Shamma, J. W. Fleshman, P. R. Wiser, and H. Versnel. Organization of response areas in ferret primary auditory cortex. *Journal of Neurophysiology*, 69:367–383, 1993.
- [133] S. A. Shamma and D. J. Klein. The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *Journal of the Acoustical Society of America*, 107(5):2631–2644, 1990.
- [134] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326, 1979.
- [135] M. Steinschneider, D. H. Reser, Y. I. Fishman, C. E. Schroeder, and J. C. Arezzo. Click train encoding in primary auditory cortex of the awake monkey: evidence for two mechanisms subserving pitch perception. *The Journal of the Acoustical Society of America*, 104:2935–2955, 1998.
- [136] K. J. Stratford, K. Tracy-Hornoch, K. A. C. Martin, N. J. Bannister, and J. J. Jack. Excitatory synaptic inputs to spiny stellate cells in cat visual cortex. *Nature*, 382:258–261, 1996.
- [137] E. Terhardt. Pitch, consonance, and harmony. *Journal of the Acoustical Society of America*, 55(5):1061–1069, 1974.
- [138] F. E. Theunissen, K. Sen, and A. J. Doupe. Spectro-temporal receptive fields of non-linear auditory neurons obtained with natural sounds. *The Journal of Neuroscience*, 20:2315–2331, 2000.

- [139] M. Tsodyks, K. Pawelzik, and H. Markram. Neural networks with dynamic synapses. *Neural Computation*, 10:821–835, 1998.
- [140] L. P. van Noorden. *Temporal coherence in the perception of tone sequences*. PhD thesis, Eindhoven University of Technology, 1975.
- [141] N. F. Viemeister and C. Plack. *Human Psychophysics*, volume 3, chapter Time analysis. Springer-Verlag, New York, 1993.
- [142] N. F. Viemeister and G. H. Wakefield. Temporal integration and multiple looks. *The Journal of the Acoustical Society of America*, 90:858–865, 1991.
- [143] D. L. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE transactions on neural networks*, 10(3):684–697, 1999.
- [144] K. Wang and S. A. Shamma. Self-normalization and noise-robustness in early auditory representations. *IEEE transactions on speech and audio processing*, 2(3):421–435, 1994.
- [145] K. Wang and S. A. Shamma. Spectral shape analysis in the central auditory system. *IEEE transactions on speech and audio processing*, 3(5):382–395, 1995.
- [146] M. Weintraub. *A theory and computational model of auditory monaural sound separation*. PhD thesis, Stanford University, 1985.
- [147] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE transactions on information theory*, 38(2):824–839, 1992.
- [148] W. A. Yost. *fundamentals of hearing: An introduction*. Academic Press, San Diego, fourth edition, 2000.

- [149] W. A. Yost, D. H. Jr Dye, and S. Sheft. A simulated "cocktail party" with up to three sound sources. *Perceptual Psychophysics*, 58(7):1026–1036, 1996.
- [150] V. Zarzoso and A. K. Nandi. *Blind Estimation Using Higher-order Statistics*, chapter Blind Source Separation, pages 167–252. Kluwer Academic Publishers, Boston, 1999.