ABSTRACT

Title of dissertation: Single-Microphone Speech Enhancement Inspired by Auditory System

Majid Mirbagheri, Doctor of Philosophy, 2014

Dissertation directed by: Professor Shihab Shamma, Department of Electrical and Computer

Enhancing quality of speech in noisy environments has been an active area of research due to the abundance of applications dealing with human voice and dependence of their performance on this quality. While original approaches in the field were mostly addressing this problem in a pure statistical framework in which the goal was to estimate speech from its sum with other independent processes (noise), during last decade, the attention of the scientific community has turned to the functionality of human auditory system. A lot of effort has been put to bridge the gap between the performance of speech processing algorithms and that of average human by borrowing the models suggested for the sound processing in the auditory system.

In this thesis, we will introduce algorithms for speech enhancement inspired by two of these models i.e. the cortical representation of sounds and the hypothesized role of temporal coherence in the auditory scene analysis. After an introduction to the auditory system and the speech enhancement framework we will first show how traditional speech enhancement technics such as wiener-filtering can benefit

on the feature extraction level from discriminatory capabilities of spectro-temporal representation of sounds in the cortex i.e. the cortical model.

We will next focus on the feature processing as opposed to the extraction stage in the speech enhancement systems by taking advantage of models hypothesized for human attention for sound segregation. We demonstrate a mask-based enhancement method in which the temporal coherence of features is used as a criterion to elicit information about their sources and more specifically to form the masks needed to suppress the noise.

Lastly, we explore how the two blocks for feature extraction and manipulation can be merged into one in a manner consistent with our knowledge about auditory system. We will do this through the use of regularized non-negative matrix factorization to optimize the feature extraction and simultaneously account for temporal dynamics to separate noise from speech.

SINGLE-MICROPHONE SPEECH ENHANCEMENT INSPIRED BY AUDITORY SYSTEM

by

Majid Mirbagheri

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
2014

Advisory Committee:
Professor Shihab Shamma, Chair/Advisor
Professor Carol Espy Wilson
Professor Yiannis Aloimonos
Professor Timothi Horiuchi
Professor Mounya Elhilali

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Shihab Shamma for giving me the freedom and motivation to work on challenging and extremely interesting projects over the past five years. He has always made himself available for help and advice. It has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to thank my PhD committee, Professor Carol Espy-Wilson, Professor Yiannis Aloimonos, Professor Ramani Duraisawami, Professor Timothy Horiuchi, and Professor Mounya Elhilali for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing the manuscript. Special thanks go to Dr Janathan Fritz for sharing his insightful views with me and encouraging me during all my years of work in neural systems laboratory.

My former and current colleagues at the neural systems laboratory have enriched my graduate life in many ways and deserve a special mention. Dr Nima Mesgarani helped me start-off by involving me in some interesting projects. Yanbo Xu and Sahar Akram deserve special thanks for all the fun discussions and collaborations we had together. My interaction with Stephen David, Bernhard Englitz, Xinhui Zhou, Kevin Donaldson, Diego Elgueda, Nik Francis, Pingbo Yin, Lakshmi Krishnan, James Snow, Kai Lu, Jenna Neckritz, and Shin Chang made all these year fruitful and enjoyable for me.

I also like to thank my best friends, Mohsen, Ehsan, Sina, Dana, Nima, Hossein and Morteza who have been there for me and inspired me in many way throughout my life. As well I feel obligated to thank my sisters, Saeedeh and Sepideh for the support and presence they never withheld from me.

Finally, my greatest thanks go to my mother and father who always stood by me and allowed me to be as ambitious as I wanted. It was under their watchful eye that I gained so much drive and an ability to tackle challenges head on. Words cannot express the gratitude I owe you.

College Park, June 2014
Majid

Table of Contents

List of Figures

List of Abbreviations

AMS	Analysis Modification Synthesis
STFT	Short Time Fourier Transform
DFT	Discrete Fourier Transform
CNS	Central Neural System
IC	Inferior Colliculus
MGN	Medial Geniculate Nucleus
ASR	Automatic Speech Recognition
SNR	Signal-to-Noise Ratio
VAD	Voice Activity Detection
HOSVD	Higher Order Singular Value Decomposition
RBF	Radial Basis Function
NSTMF	Nonlinear Spectro-Temporal Modulation Filtering
STMF	Spectro-Temporal Modulation Filtering
MOS	Mean Opinion Score
ASA	Auditory Scene Analysis
AISWR	Auditory-Inspired Spectral Weighting Rule
MI	Mutual Information
KNN	K-Nearest Neighbor
MMSE	Minimum Mean Square Error
BT	Block Thresholding
EMBSD	Enhanced Modified Bark Spectral Distortion
STCC	Short Time Coherence Coefficient
NMF	Nonnegative Matrix Factorization
CNMF	Convolutive Nonnegative Matrix Factorization
PESQ	Perceptual Evaluation of Speech Quality

Chapter 1: Introduction

Speech processing applications have gained plenty of interest during last decade as the machine-human interaction through speech enters daily lives of people more and more. Voice controlled devices, smart phone applications and automated customer services are just a few examples in this new wave. Naturally, by the increase in popularity the demand for more robust applications which can work anywhere and at any time also increases over time. More specifically, these applications should now detect and track a target source (Speech) of interest in the presence of acoustical disturbances such as traffic noise, back ground music or even another competing speaker. Almost always the performance of these applications is severely affected if the noise is not handled correctly.

Speech enhancement as a popular solution aims to process the noisy speech signal and to reduce the impact of the noise and enhance the sound quality i.e. listener comfort or speech intelligibility. Speech enhancement can be done using single-microphone (monaural) or multi-microphone methods. In terms of performance, single-microphone methods often fall behind multi-microphone methods, but they are usually preferred when there exist limitations in size, computational complexity or power usage. Moreover many times single-channel methods are used in multi-

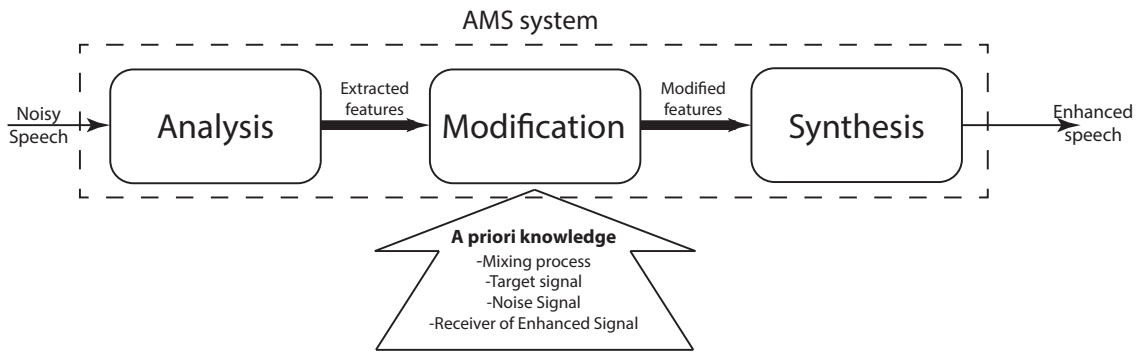

Figure 1.1: General single-channel speech enhancement system.

microphone systems as a post-processing stage following a beamformer [1].

Single-microphone speech enhancement has been an active area of research for over 30 years, resulting in numerous methods and algorithms. Despite the varieties of these systems, they often have a similar general structure i.e. the analysis-modification-synthesis (AMS) arrangement. Figure 1.1 demonstrates the block diagram of an AMS system.

The first two blocks exist in any enhancement system. The analysis block is responsible for extracting features that can well represent the dynamic behavior of speech. A very common choice in the existing systems is short-time Fourier transform (STFT) computed by discrete Fourier transform (DFT). Besides historical reasons, the efficient computation of STFT and its straightforward link to physical properties such as frequency content of the incoming signals have been the key factors in this choice. The greatest variety of enhancement systems arises in the modification stage where the extracted features are modified in a way that they can represent the clean speech signal. Looking at the different approaches in the

field, they can be categorized into two main groups i.e., *top-down*, and bottom-up methods [2].

In top-down approaches such as [3, 4], generative models are used to capture the statistics of features of isolated signals, as well as the effect on the features of mixing two signals. Taking advantage of the a prior knowledge about the speech, noise and the mixing process, the inference seeks the speech and noise signals that are most likely given the observed mixture. The decomposition of the spectrogram (or other time-frequency representation) into its constituent sources emerges as a by-product of this inference.

In bottom-up approaches, segmentation rules operate on low-level features to determine which regions of the mixture representation belong to the target speech. Often times in these methods a measure of target speech dominance is estimated for each feature and used to modify the representation.

Although original enhancement methods were mostly addressing this problem in a pure signal processing framework, during last decade numerous approaches specially in the bottom-up class have been proposed, inspired by the functionality of human auditory system [5,6]. Some of these methods have tried to integrate computational models suggested for the sound processing in the auditory system in the design of analysis-modification-synthesis blocks, while others have taken into account our knowledge about hearing.

In this thesis, we follow this trend by presenting speech enhancement methods in both categories that take advantage of auditory models in the design of analysis and modification blocks. The dissertation is organized in six chapters. Following

this introduction, we present an overview of the organizational structure of auditory pathway, starting from the external ear and ending in the primary auditory cortex. We also present the computational model suggested for sound processing in peripheral and central auditory system.

In chapter 3, we will show how traditional speech enhancement methods can benefit on the feature extraction level from discriminatory capabilities of spectrotemporal representation of sounds in the cortex i.e. the cortical model. We present a method that identifies nonspeech segments of the noisy signal uses them to compute the transformations needed in modification stage all performed in cortical domain.

In chapter 4, we put the feature extraction aside for a while and focus on the modification stage. We overview the coherence-based model for auditory scene analysis and see how it can serve as a foundation for speech enhancement. We present a button-up mask-based enhancement method that uses mutual information as a measure of coherence between features and a cue signal representing the target source to form gain functions needed in modification stage. We provide examples in which loudness and estimated pitch are used to clean noisy speech signals.

Chapter 5 explores how the spectroteporal feature extractors can adapt themselves for better separation of noise from speech by merging the analysis and modification stages. We present a noise reduction scheme based on regularized Non-negative Matrix Factorization (NMF) in which the feature extractors (atom) simultaneously adapt and take part in the separation process. In order to demonstrate the effectiveness of the proposed methods, we provide performance comparison results for all three methods.

Finally we conclude in Chapter 6 with an overview of the contributions presented in this work and discuss the future works for auditory inspired speech enhancement.

Chapter 2: Auditory System

2.1 The Auditory Pathway

Hearing in humans and other vertebrates is handled by the auditory system. This system provide means to capture information about the surrounding objects through the sounds they generate. The sound itself is the result of the propagating energy produced by vibrating objects in an elastic medium in the form of a disturbance or pressure wave. The ear as the peripheral input gate of the auditory system receive these vibrations and transduce the mechanical energy into electro-chemical signals in the nervous system. At the core of the system, brain will process these signals and extract certain attributes of the sound source such as location, content and identity. in this chapter we will briefly review what we know about the auditory system and see how sound is processed and perceived all the way from the external ear to regions in central nervous system (CNS).

2.1.1 Ear

In order to hear sounds, ear is responsible for capturing the mechanical energy (sound), transmitting it to the ears's receptive organ and transducing it into electri-

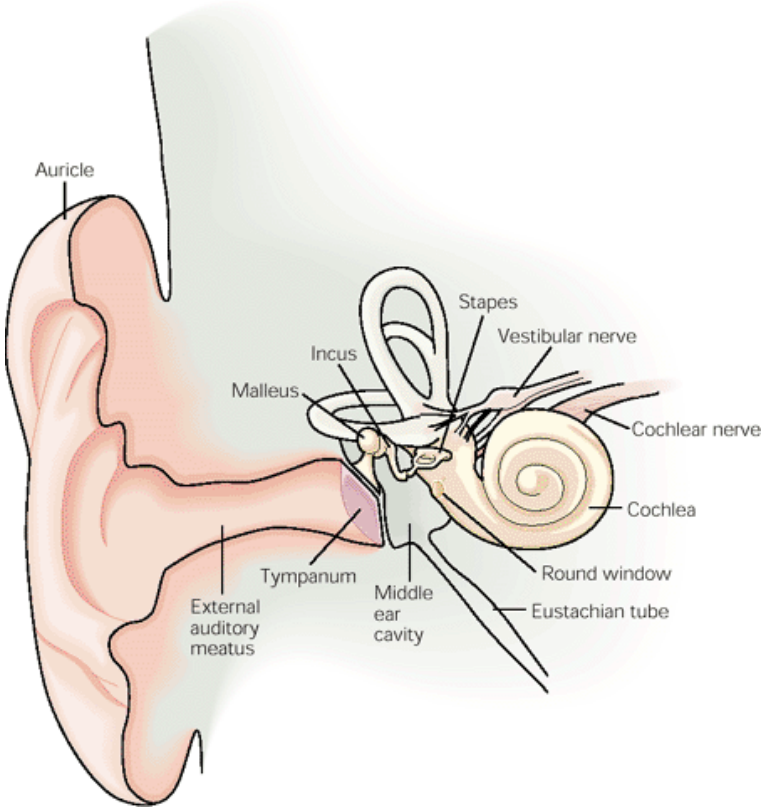

Figure 2.1: The structure of the human ear. (Adapted from Noback 1967)

cal signals that can be analyzed by the nervous system. These tasks are respectively accomplished by the three functional parts of ear i.e. external ear, the middle ear and the internal ear [7] .Figure 2.1 illustrates the structure of the ear. The external ear, especially the prominent auricle, focuses sound into the external auditory meatus. Alternating increases and decreases in air pressure vibrate the tympanum. These vibrations are conveyed across the air-filled middle ear by three tiny, lined bones: the malleus, the incus, and the stapes. Vibration of the stapes stimulates the cochlea, the hearing organ of the inner ear.

The cochlea shown in figure 2.2 in the inner ear consists of three fluid-filled

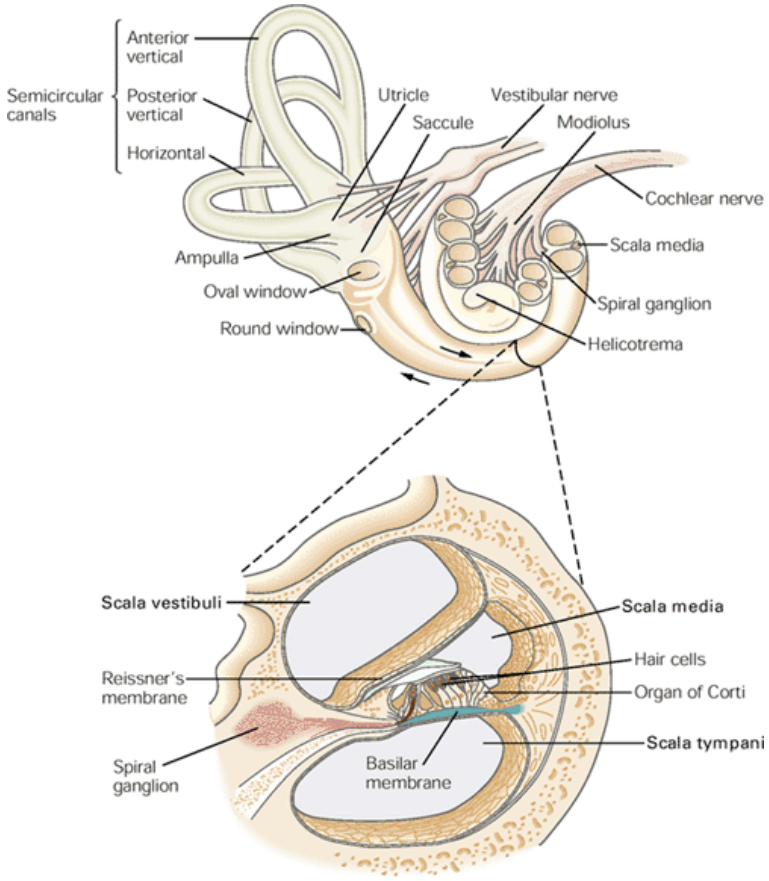

Figure 2.2: The Cochlea structure

compartments throughout its entire length of 33 mm. A cross section of the cochlea shows the arrangement of the three ducts. The oval window, against which the stapes pushes in response to sound, communicates with the scala vestibuli. The scala tympani is closed at its base by the round window, a tick, flexible membrane. Between these two compartments lies the scala media, an endolymph-filled tube whose epithelial lining includes the 16,000 hair cells surrounding the basilar membrane.

2.1.2 Functional anatomy of the cochlea

Illustrated in figure 2.3, the basilar membrane is a mechanical analyzer of sound frequency. The mechanical properties of the basilar membrane are key to the cochlea's operation. In brief, the membrane is tapered and it is stiffer at one end than at the other. The dispersion of fluid waves causes sound input of a certain frequency to vibrate some locations of the membrane more than the other locations. As shown in experiments by Nobel Prize laureate George von Bekesy, high frequencies lead to maximum vibrations at the basal end of the cochlear coil (narrow, stiff membrane), and low frequencies lead to maximum vibrations at the apical end of the cochlear coil (wide, more compliant membrane).

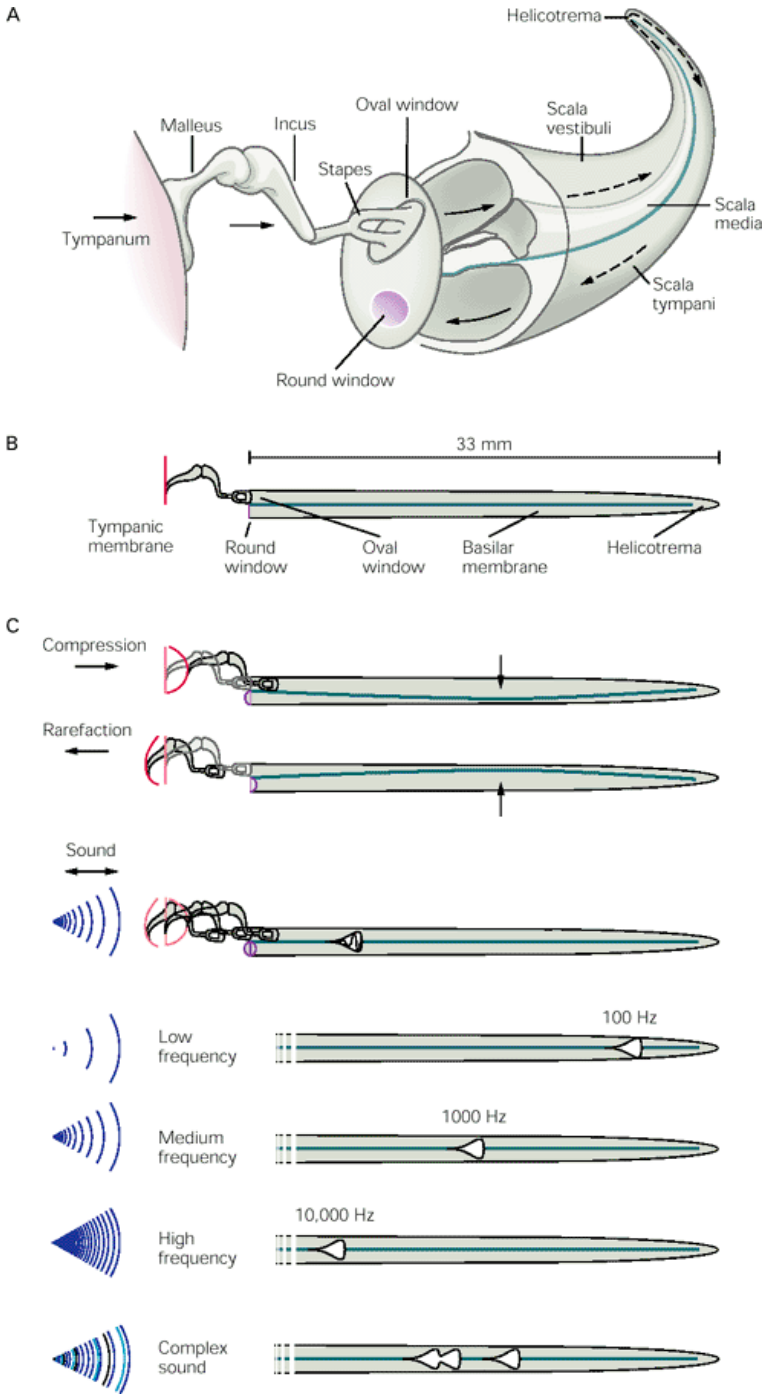

Figure 2.3: The basilar membrane and its frequency analysis mechanism

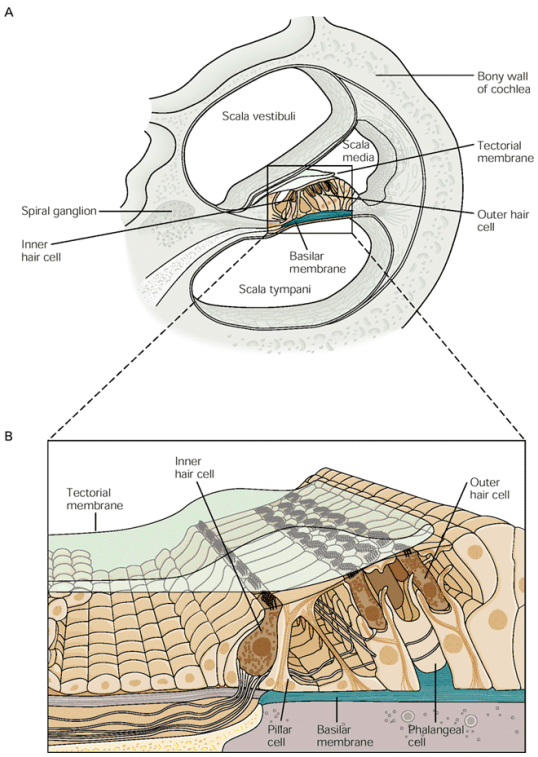

Figure 2.4: Cellular architecture of the organ of Corti in the human cochlea

2.1.3 Cellular architecture of the organ of Corti

The organ of Corti shown in figure 2.4 is the organ in the inner ear of mammals that contains auditory sensory cells, or hair cells. The organ contains some 16,000 hair cells arrayed in four rows: a single row of inner hair cells and three of outer hair cells. The mechanically sensitive hair bundles of these receptor cells protrude into endolymph, the fluid contents of the scala media. The hair bundles of outer hair cells are attached at their tops to the lower surface of the tectorial membrane, a gelatinous shelf that extends the full length of the basilar membrane. The basic architecture of the organ of Corti is similar for all mammals.

Hair cells in the cochlea are stimulated when the basilar membrane is driven up and down by differences in the fluid pressure between the scala vestibuli and scala tympani. Because this motion is accompanied by shearing motion between the tectorial membrane and organ of Corti, the hair bundles that link the two are deflected. This deflection initiates mechanoelectrical transduction of the stimulus. When the basilar membrane is driven upward, shear between the hair cells and the tectorial membrane deflects hair bundles in the excitatory direction, toward their tall edge. At the midpoint of an oscillation the hair bundles resume their resting position. When the basilar membrane moves downward, the hair bundles are driven in the inhibitory direction (Figure 2.5).

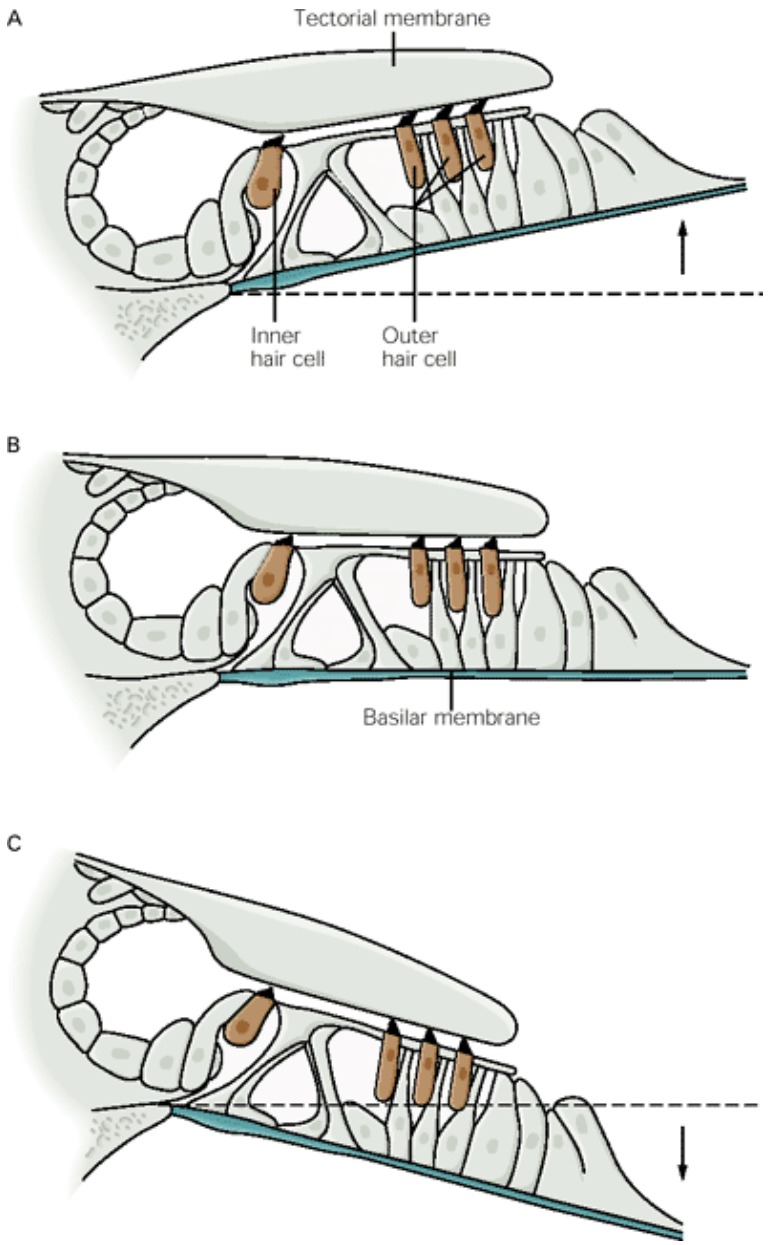

Figure 2.5: Hair cell stimulation by basilar membrane stimulation

The receptor potential in mammalian outer hair cells triggers active vibrations of the cell body (figure 2.6). Mammals have not improved hearing sensitivity, but the outer hair cells evolved only in them. As a result, they have extended the hearing

range and frequency selectivity which is of particular benefit for humans, because it enables sophisticated speech and music.

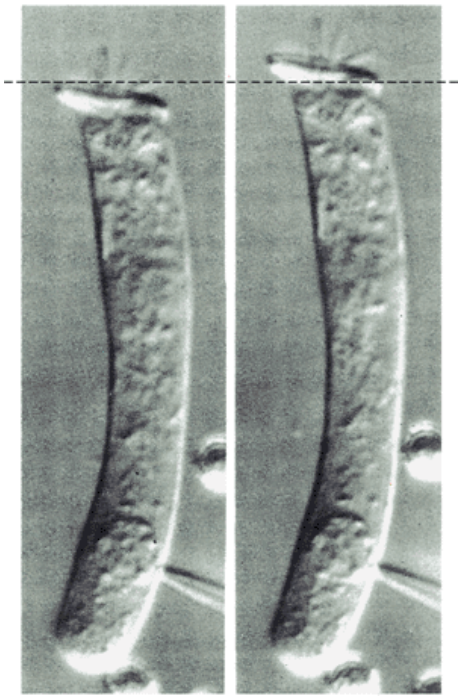

Figure 2.6: Outer hair cell

2.1.4 Structure of inner hair cells

As shown in figure 2.7 The cylindrical hair cell is joined to the adjacent supporting cells by a junctional complex around its apical perimeter. From the cells apical surface extends the hair bundle, the mechanically sensitive organelle. Afferent and efferent synapses occur upon the basolateral surface of the plasma membrane. The bundle comprises some 60 stereocilia, each a cylinder with a tapered base, arranged in stepped rows of varying length. Deflection of the hair bundle to the right, the positive stimulus direction, depolarizes the hair cell; movement in the opposite

direction elicits a hyperpolarization.

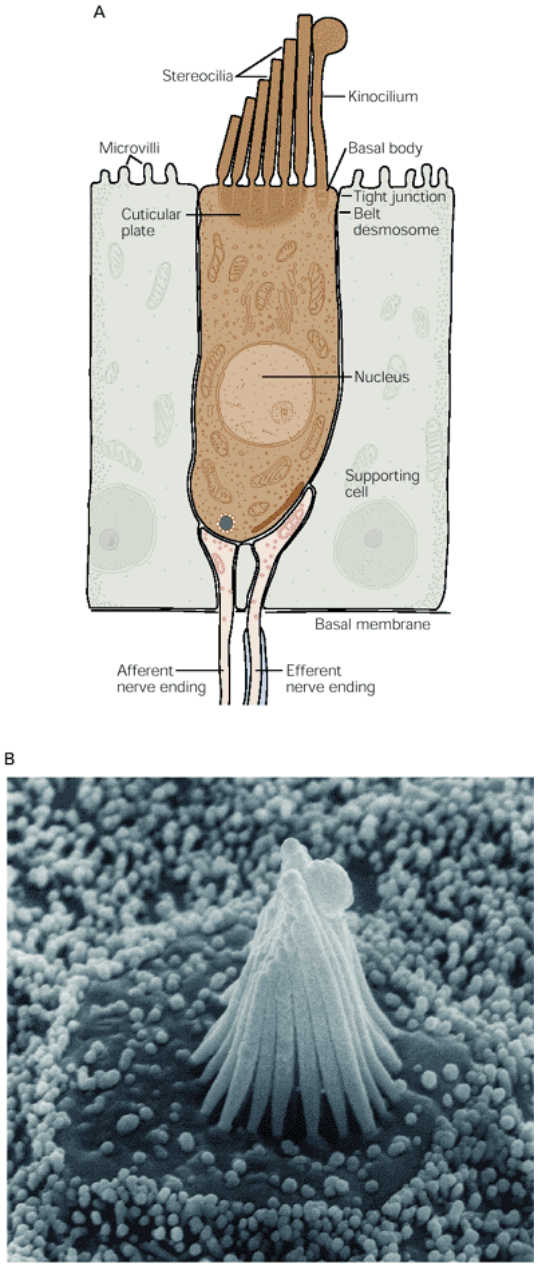

Figure 2.7: Structure of a vertebrate hair cell

2.1.5 Transformation of mechanical energy into neural signals

Deflection of the hair bundle initiates mechanoelectrical transduction. This involves a mechanism for gating of ion channels that is fundamentally different from those employed in such electrical signals as the action potential or postsynaptic potential. The opening and closing of transduction channels is regulated by the tension in the elastic structure within the hair bundle. Figure 2.8 illustrates this mechanism. The ion channels that participate in mechanoelectrical transduction in hair cells are gated by elastic structures in the hair bundle. The channel is assumed to be a membrane-spanning protein with a cation- selective pore. When the hair bundle is at rest, each transduction channel clatters between closed and open states, spending most of its time shut. Displacement of the bundle in the positive direction increases the tension in the gating spring, here assume to be a tip link attached to each channel's molecular gate. The enhanced tension promotes channel opening and the influx of cations, thereby producing a depolarizing receptor potential.

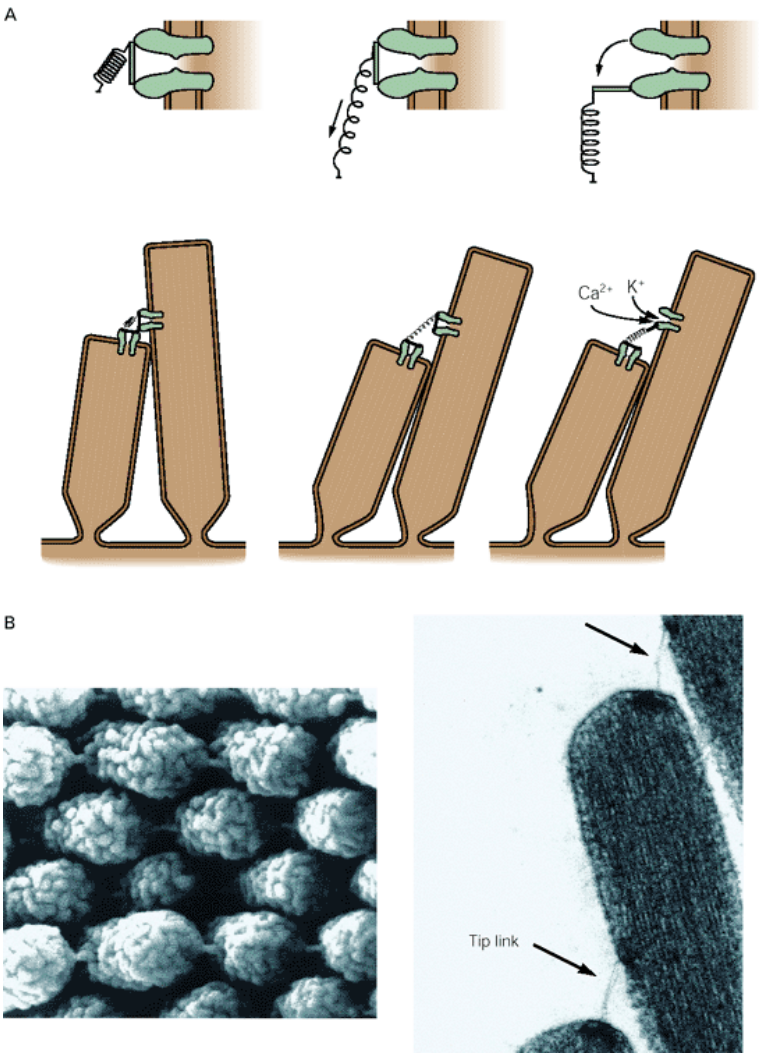

Figure 2.8: A model for the mechanism of mechanoelectrical transduction by hair cells

2.1.6 Innervation of the organ of Corti

The great majority of afferent axons end on inner hair cells, each of which constitutes the sole terminus for an average of 10 axons. A few afferent axons of small caliber provide diffuse innervation to the outer hair cells. Efferent axons largely innervate outer hair cells, and do so directly. In contrast, efferent innervation

of inner hair cells is sparse and is predominantly axoaxonic, at the ending of afferent nerve fibers. An illustration of the innervation is shown in figure 2.9.

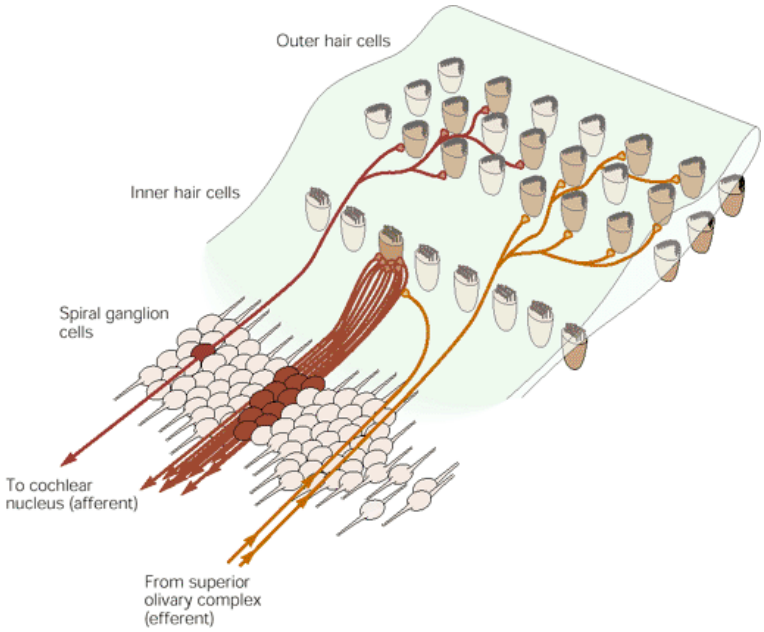

Figure 2.9: Innervation of the organ of Corti.

2.1.7 Computational model for peripheral auditory processing

Computational Models aim to mimc the functionality of systems. For the peripheral auditory processing numerous computational models have been suggested based on neurophysiological data gathered from mammalians peripheral stage of processing [8,9]. The specific model we will use though this thesis stablished by Wang *et al* [10] was preferred over others for the its biological foundation and perceptual relevance which have been shown through analytical and experimental investigations. Throughout this section, we will discuss how an *auditory spectrogram* is computed based the the original work in [8].

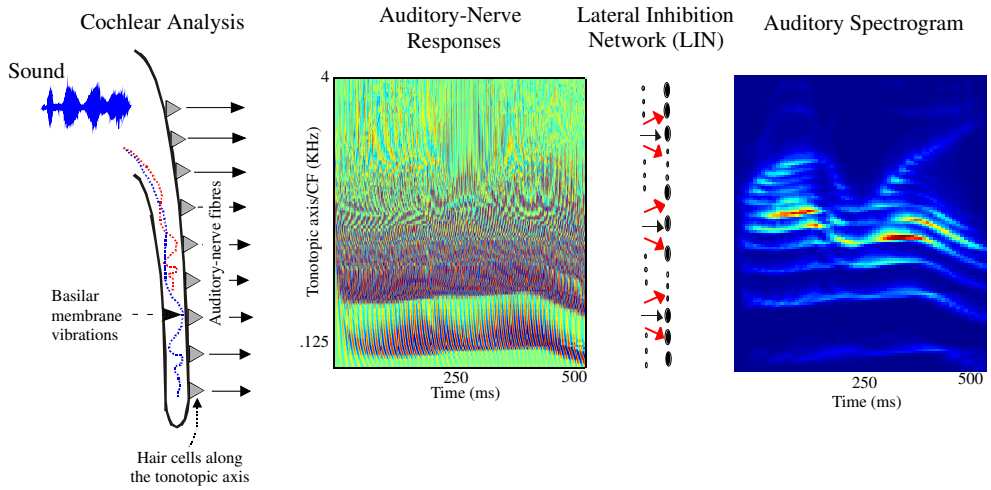

Figure 2.10: Schatic of peripheral auditory processing modeled as a three step process.

The computation involves a stage of wavelet analysis followed by a series of linear and nonlinear transformations applied on the acoustic waveform. Figure 2.10 shows a schematic of the peripheral auditory processing i.e. the computation of the auditory spectrogram. The model can be formulized through the following three steps computation:

$$y_1(t,x) = s(t) *_t h(t;x) \tag{2.1}$$

$$y_2(t,x) = g(\delta_t y_1(t,x)) *_t \omega(t) \tag{2.2}$$

$$y_3(t,x) = max(\delta_x y_2(t,x), 0) *_t \mu(t;\tau) \tag{2.3}$$

With $s(t)$ being the acoustic waveform, equation (2.1) models the frequency analysis mechanism of cochlea as a filter bank consisting of constant-Q ($Q = 4$)

highly asymmetric bandpass filters, $h(t, x)$ that are uniformly spread over the frequency axis. The filters span a 5.3 octave range on the frequency axis with 24 filters in each octave.

The next stage (equation (2.2)) models the conversion of the basilar membrane outputs into inner hair cell intra-cellular potentials i.e. $y_2(t, x)$. The conversion involves the following operations: a high-pass filtering (the fluid-cilia coupling), a nonlinear compression (gated ionic channels) denoted by a nonlinear function $g(\cdot)$, and a low-pass filtering by the filter $\omega(\cdot)$ (hair cell membrane leakage).

The final step (equation (2.3)) mimics the functionality of the the lateral inhibitory network that detects discontinuities in the responses across the tonotopic axis of the auditory nerve array and the sharpening of the filter-bank frequency selectivity observed in the cochlear nucleus. 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 the time constant τ. This stage effectively sharpens the bandwidths of the cochlear filters from about $Q = 4$ to 12, as explained in detail in [10].

Overall the resulting spectrogram acts as a temporal envelope tracker for the components interacting with each other within the bandwidths of each filter.

2.1.8 The central auditory pathway

The central auditory pathways extend from the cochlear nucleus to the auditory cortex. Postsynaptic neurons in the cochlear nucleus send their axons to other

centers in the brain via three main pathways: the dorsal acoustic stria, the intermediate acoustic stria, and the trapezoid body. The first binaural interactions occur in the superior olivary nucleus, which receives input via the trapezoid body. In particular, the medial and lateral divisions of the superior olivary nucleus, along with axons from the cochlear nuclei, project to the inferior colliculus in the midbrain via the lateral lemniscus. Each lateral lemniscus contains axons relaying input from both ears. Cells in the colliculus send their axons to the medial geniculate nucleus of the thalamus. The geniculate axons terminate in the primary auditory cortex, a part of the superior temporal gyrus (Figure 10). Information flows from cochlear hair cell to neurons whose cell bodies lie in the cochlear ganglion. The pattern of afferent innervations in the human cochlea emphasizes the functional distinction between inner and outer hair cells. At least 90% of the cochlear ganglion cells terminate on inner hair cells. Each axon innervates only a single hair cell, but each inner hair cell directs its output to several nerve fibers, on average nearly 10. The output of each inner hair cell is sampled by many nerve fibers, which independently encode information about the frequency and intensity of sound. The tonotopic organization of the auditory neural pathways begins at the earliest possible site, immediately postsynaptic to inner hair cells.

The acoustical sensitivity of axons in the cochlear nerve mirrors the innervation pattern of spiral ganglion cells. Each axon is most responsive to stimulation at a particular frequency of sound, its characteristic frequency. Stimuli of lower or higher frequency also evoke responses, but only when presented at greater intensities. The relation between sound-pressure level and firing rate in each fiber of the cochlear

nerve is approximately linear. Difference in neuronal responsiveness originate at the synapses between inner hair cells and afferent nerve fibers. Nerve terminals on the surface of a hair cell nearest the axis of the cochlear spiral belong to the afferent neurons of lowest sensitivity and spontaneous activity. The multiple innervations of each inner hair cell are therefore not completely redundant. Instead, because of systematic differences in the rate of transmitter release or in postsynaptic responsiveness (or both), the output from a given hair cell is directed into several parallel channels of differing sensitivity and dynamic range.

Three important general principles emerge from connections in the brain stem. First, acoustical information is processed in parallel pathways, each of which is dedicated to the analysis of a particular feature of auditory information. Second, the various cell types of the cochlear nuclei project to specific relay nuclei, so that the separation of information streams commence within the cochlear nuclei. Finally, there is extensive interaction between auditory structures on the two sides of the brain stem. The medial superior olive performs a specific function in a readily intelligible way. The ability to localize sound sources along the azimuthal axis stems in part from the processing of information about auditory delays.

The inferior colliculus (IC) is divisible into two major components. Because it contains many neurons sensitive to interaural timing or intensity differences, the IC is apparently involved in sound localization. The medial geniculate body (MGN) constitutes the thalamic relay of the auditory system. This nuclear complex comprises at least three subdivisions of which the principal nucleus is the best understood. Most neurons in MGN are sharply tuned to specific stimulus frequencies, and most

are responsive to stimulation through either ear.

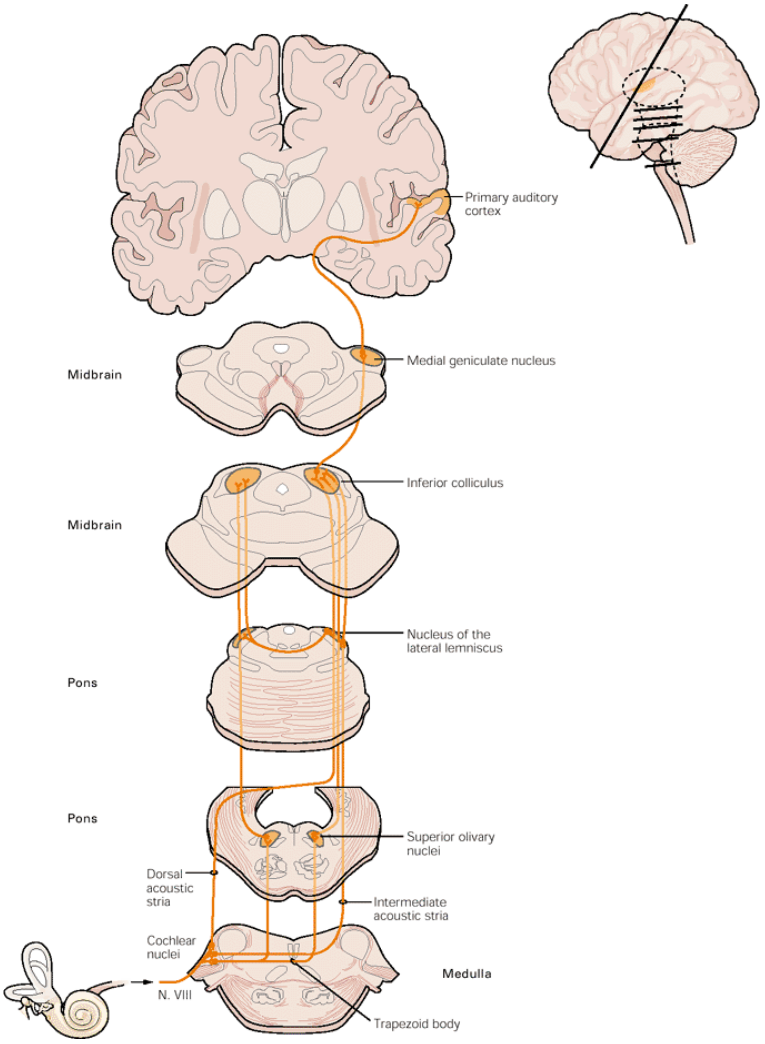

Figure 2.11: The central auditory pathway

The ascending auditory pathway terminates in the cerebral cortex, where several distinct auditory areas occur on the dorsal surface of the temporal lobe. The most prominent projection from the ventral nucleus of the MGN extends to the primary auditory cortex (A1).

It should be pointed out that due to the anatomical complexity of the pathways, the neural morphology of cells and circuitry, and the unknown nature of the neural code, our understanding of the structure and function of the central auditory nervous system is far less than that of the periphery. However the brain imaging besides psycho-acoustical and neurophysiological studies have vastly broaden our knowledge and provided us with tools to gain insight toward the function of the central auditory system and the processes in the brain for sound perception.

2.1.9 Computational Model

There is no consensus regarding the real role of the cortical circuitry in sound perception [11], but a *simplistic* view about the neurons in the cortex is that they serve as "feature extractors" for the processing of sound. Aligned to this view, the model we describe and use in this thesis is proposed by Chi *e*t al. [12]. They derived the model based on the physiological detain animals [13–15], and psycho-acoustical data in humans [16].

The model consists of a multi-scale filter-bank represented by impulse responses in the form of spectrotemporal Gabor functions [16].

Each of these two-dimensional filters are tuned to a range of temporal (denoted ω, or rate) and spectral (denoted Ω, or scale) modulations. The overall impulse response of each filter is a "separable" spectrotemporal modulation function $\mathcal{RF}$ which can be computed as the product of two marginal functions i.e. a spatial

impulse response $h_{\mathcal{RF}}(x;\Omega_c,\phi_c)$ and temporal impulse response $g_{\mathcal{RF}}(t;\omega_c,\theta_c)$ (as shown in the Figure 2.12) mathematically formulated as:

$$\begin{aligned} g_{\mathcal{RF}}(t;\omega_c,\theta_c) &= g(t;\omega_c)\cos\theta_c + \hat{g}(t;\omega_c)\sin\theta_c \\ h_{\mathcal{RF}}(x;\Omega_c,\phi_c) &= h(x;\Omega_c)\cos\phi_c + \hat{h}(t;\Omega_c)\sin\phi_c \\ \mathcal{RF}(t,x;\omega_c,\theta_c,\Omega_c,\phi_c) &= g_{\mathcal{RF}}(t;\omega_c,\theta_c)\cdot h_{\mathcal{RF}}(x;\Omega_c,\phi_c) \end{aligned} \tag{2.4}$$

The parameters in the models determine the selectivity of cortical neurons to spectral local shapes, rate movements of spectra, as well as direction of movement (upward or downward). In this way the spectrotemporal response of the neuron to an input spectrogram $y(t,x)$ can be computed as:

$$\begin{aligned} r(t,x;\omega_c,\theta_c,\Omega_c,\phi_c) &= y(t,x) *_{xt} \mathcal{RF}(t,x;\omega_c,\theta_c,\Omega_c,\phi_c) \\ &= y(t,x) *_{xt} [g_{\mathcal{RF}}(t;\omega_c,\theta_c)\cdot h_{\mathcal{RF}}(x;\Omega_c,\phi_c)] \\ &= y(t,x) *_{xt} [g\cdot h\cos\theta_c\cos\phi_c + g\cdot\hat{h}\cos\theta_c\sin\phi_c \\ &\quad + \hat{g}\cdot h\sin\theta_c\cos\phi_c + \hat{g}\cdot\hat{h}\sin\theta_c\sin\phi_c] \end{aligned} \tag{2.5}$$

In the next chapter we use this model in the complex form in which the output is reduced to a 4 dimensional complex-valued mapping $r(t,x;\omega,\Omega)$ 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 [17]. A schematic of the multi-scale wavelet analysis performed by cortical neurons is shown in Figure 2.12. It shows how the input spectrogram is

decomposed through the various filters into a four-dimensional complex-valued response (time, frequency, rate, and scale). The right panel in this figure shows the magnitude response of 4 different modulation selective filters. Fast temporal envelopes in the original speech corresponding to rates +32Hz and -32Hz are detected by the two fast filters while the 8Hz filter capture the slower envelope dynamics representing the overall patterns in the input spectrogram. The upward vs. downward filters capture different patterns in the input representing the orientation selectivity of neurons in the cortical model.

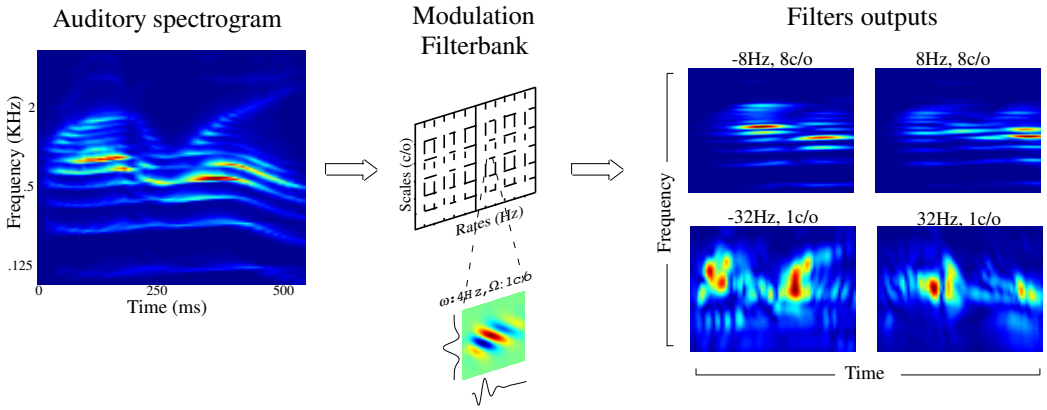

Figure 2.12: The cortical multi-scale representation of sound.

Chapter 3: Nonlinear Filtering of Spectro-Temporal Modulations for Speech Enhancement

3.1 Overview

Noise suppression for speech applications is used to enhance the perceptual quality of speech or to improve the performance of speech processing and communication systems. It can also play an important role in automatic speech recognition systems (ASR) by improving their robustness in noisy environments. This has been an active area of research for over fifty years, mostly framed as a statistical estimation problem in which the goal is to estimate speech from its sum with other independent processes (noise). This strategy requires an underlying statistical model of the signal and noise, as well as an optimization criterion. In some of the earliest work, one approach was to estimate the speech signal itself [18]. When the objective is expressed as minimization of mean-square error, the problem reduces to the design of an optimum Wiener filter. Estimation can also be achieved in the frequency domain, as in methods such as the spectral subtraction [18], the signal subspace approach [19], and the estimation of the short-term spectral magnitude [20]. Estimation in the frequency domain is superior to the time domain as it offers better

initial separation of the speech from noise, which in turn (1) results in easier implementation of optimal/heuristic approaches, (2) simplifies the statistical models because of the decorrelation of the spectral components, and (3) facilitates integration of psychoacoustic models [21].

Recent psychoacoustic and physiological findings in mammalian auditory systems, however, suggest that the spectral decomposition is only the first stage of several further transformations in the representation of sound. Specifically, it is thought that neurons in the auditory cortex decompose the spectrogram further into its spectro-temporal modulation content [22]. This finding has inspired a multi-scale model representation of speech modulations that has proven useful in assessment of speech intelligibility [23], discriminating speech from nonspeech signals [24], and in accounting for a variety of psychoacoustic phenomena [25]. A key feature of this analysis is that extracted modulations of noise and speech often have a very different character, and hence their representations are well separated making it readily suitable in the context of speech enhancement applications. Filtering of such spectro-temporal modulations has already been demonstrated in the enhancement of speech quality in [26]. In that work, the representation of noise in the "modulation domain" is first estimated, and then used to construct denoising filters to remove it from the speech signal.

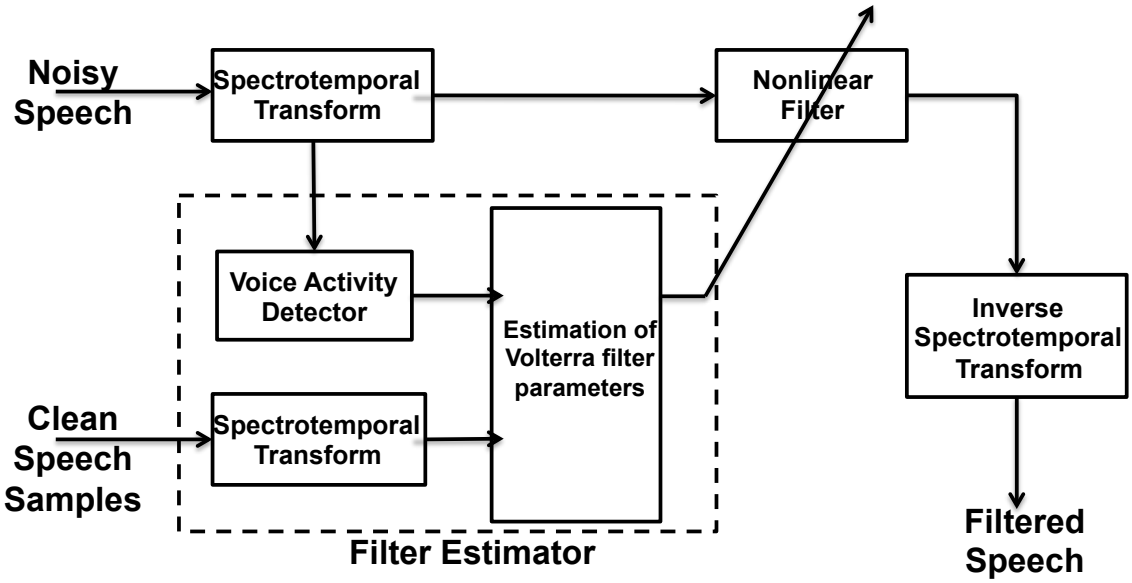

Figure 3.1: Schematic of the proposed nonlinear filtering of spectro-temporal modulations

In this chapter, we offer a new approach as introduced in [27] that differs from [26] in two major ways: (1) In addition to modeling the spectro-temporal representation of noise, we utilize the statistics of clean speech in the estimation of denoising filters; (2) we also take into account the dynamics of speech by using

nonlinear filters [28]. As a result, the quality of the filtered speech is far better preserved while reducing the background noise. A key component of this approach is the invertible auditory cortical model, which can be used to transform the noisy signal, and subsequently invert it back to the acoustic signal once nonlinear filtering is applied. Figure 3.1 illustrates the diagram of the proposed method. Details of each stage are provided in the following sections.

3.2 Spectro-temporal modulation analysis

The auditory model was inspired by psychoacoustic and neurophysiological findings in the early and central stages of the auditory pathway. The early stage converts the sound waveform into an *auditory spectrogram* - roughly akin to a time-frequency distribution along a tonotopic (logarithmic frequency) axis [17]. The second (cortical) stage performs a two-dimensional wavelet transform of the auditory spectrogram, thus providing an estimate of its spectral and temporal modulation content. It is computationally implemented by a bank of two-dimensional (spectro-temporal) filters that are selective to different modulation parameters ranging from slow to fast *rates* temporally and narrow to broad *scales* spectrally.

The spectro-temporal impulse responses (or "receptive fields") of these filters are centered at different frequencies along the tonotopic axis. Therefore, the basic mathematical formulation of the model can be summarized as follows:

$$r(t, f; \omega, \Omega) = y(t, f) *_{tf} h(t, f; \omega, \Omega) \tag{3.1}$$

where $y(t, f)$ is the auditory spectrogram, $h(/dot)$ the spectrotemporal impulse response, and $r(t, f; \omega, \Omega)$ the rate-scale representation. Since the cortical stage (Equation (3.1)) is linear and invertible, we can readily reconstruct the auditory spectrogram $y(t, f)$ from its modified rate-scale representation, $\hat{r}(t, f; \omega, \Omega)$. The reconstruction of an audio waveform from the auditory spectrogram is achieved by an iterative method based on a convex projection algorithm described in [9]. The central stage processing is illustrated with an example in Figure 3.2.

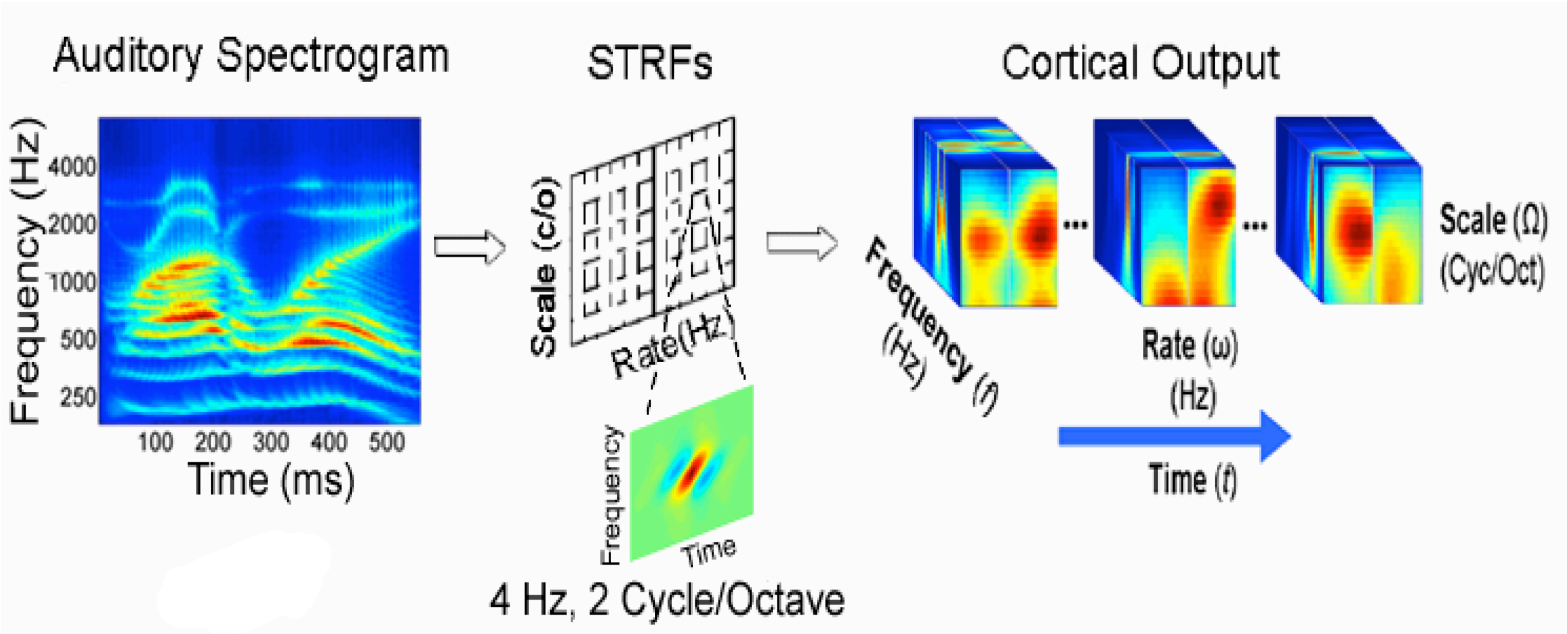

Figure 3.2: The auditory spectrogram (left) is decomposed into its spectrotemporal components using a bank of spectrotemporally selective filters. The impulse response (spectrotemporal receptive fields or STRF) of such a filter is shown in the center panel.

3.3 Feature Modification

The feature modification stage takes cortical representation of the noisy speech as input and outputs the enhanced spectro-temporal features. This modification has two step is done in two steps: first we form the second order Volterra series expansion of spectro-temporal representation of the noisy speech. Next, a linear mapping is found from the Volterra expansion of noisy to clean signal representation in spectro-temporal domain. To estimate the mappings, we first detect the non-speech segments of the noisy signal. We then generate noisy speech samples by adding the noise-only segments to stored clean samples at the estimated SNR. Having the noisy and original clean speech exemplars, we then estimate the nonlinear optimum filter from noisy to clean representations as described below. These filters can be updated as frequently as needed to track the changes in the statistics or type of the background noise.

3.3.1 Extracting Noise-only Segments

Estimation of the background noise modulations is the first step in single-microphone speech quality enhancement. This task is particularly challenging in adverse environments with low signal-to-noise ratios (SNR) and highly non-stationary background noise. Most of the proposed techniques are based on three assumptions: (1) speech and noise are statistically independent, (2) speech is not always present, and (3) the noise is more stationary than speech [21]. All such methods must employ a Voice Activity Detector (VAD), or tracking of spectral minima [21].

We used a VAD that is based on the same multi-scale spectro-temporal modulations as described in [29].

This method performs speech detection in cortical domain in two step:

1. Dimension reduction: The typical size of cortical representation is usually very large (around 7500 coefficients per frame), but the elements are highly correlated making it possible to reduce the dimension significantly using a comprehensive data set, and finding new multilinear and mutually orthogonal principal axes that approximate the subspace of smaller dimensionality spanned by these data. The training set we used consisted 1500 cortical frames from both speech and nonspeech classes. By stacking all these frame we formed a 4-D tensor D of size $5 \times 12 \times 128 \times 1500$. Using a higher order SVD (HOSVD) decomposition described in [30] we decamposed D to its *mode-n* singular vectors:

$$D = S \times_1 U_{\text{frequency}} \times_2 U_{\text{rate}} \times_3 U_{\text{scales}} \times_4 U_{\text{samples}} \tag{3.2}$$

where $U_{\text{frequency}}$, U_{rate}, and U_{scale} are orthonormal ordered matrices containing subspaces singular vectors, obtained by unfolding D along its corresponding modes. Tensor S is the core tensor with the same dimension as D. Singular matrices are then truncated so that only a desired number of principal axes are retained. As shown in Figure 3.3, in order to reduce the dimension of new sound samples in cortical domain represented by 4-D tensor A will be

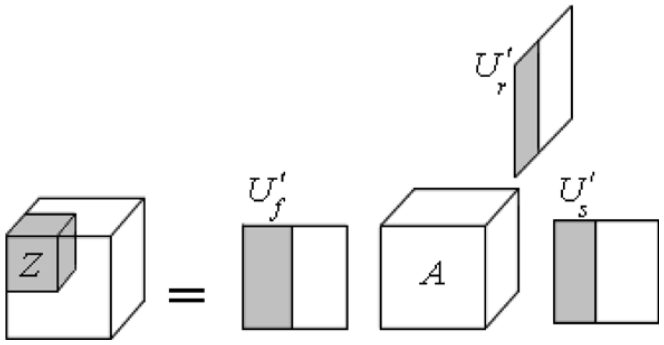

Figure 3.3: Dimensionality reduction using HOSVD

projected onto these truncated orthonormal axes $U'_{\text{frequency}}$, U'_{rate}, and U'_{scale} as following:

$$Z = A \times_1 U'_{\text{frequency}} \times_2 U'_{\text{rate}} \times_3 U'_{\text{scales}} \tag{3.3}$$

2. Classification: Speech frames are distinguished from nonspeech ones using a support vector machine (SVD) [31, 32] classifier. The optimal boundary separating the two classes is found by SVMs in such a way as to maximize the margin between separating boundary and closest samples to it (support vectors). We used the same data set used in dimension reduction stage to train SVMs with a radial basis function (RBF) kernel.

This method detects speech reliably at low SNRs (e.g., -5 dB), relying primarily on the fact that average spectro-temporal modulations of clean speech are distinctive and can be reliably detected and discriminated from other non-speech sounds [23,24].

Once noise-only segments are extracted, they can be added to the stored clean speech samples to create noisy speech signals with known clean samples.

3.3.2 Estimation of the Nonlinear Filter Parameters

The nonlinear filter can be viewed as a *nonlinear mapping* between the spectro-temporal representation of the noisy signal to its corresponding clean representation. The mapping is found in the adaptive estimation stage by learning the optimal transformation between representation of the constructed noisy signals and the corresponding clean samples. Starting from the 4-D spectro-temporal representation of noisy speech $r_n(t, f; \omega, \Omega)$, we use Volterra expansion to form a new representation for each frequency channel (f_c):

$$\begin{aligned} R_n(t, f_c) =& [1, r_n(t, f_c; \omega_1, \Omega_1), \ldots, r_n(t, f_c; \omega_{n_r}, \Omega_{n_s}) \\ & , r_n(t, f_c; \omega_1, \Omega_1)^2, r_n(t, f_c; \omega_1, \Omega_1) r_n(t, f_c; \omega_2, \Omega_2), \ldots, r_n(t, f_c; \omega_{n_r}, \Omega_{n_s})^2] \end{aligned} \tag{3.4}$$

where n_r and n_s are respectively the number of rates and scales in spectro-temporal representation. Assuming the clean version of the spectro-temporal representation for the noisy sample exists, $r_c(t, f_c; \omega_i, \Omega_j)$, the goal is then to estimate a signal that minimizes the mean squared distance to the original sample:

$$\min \sum_t (r_c(t, f_c; \omega_i, \Omega_j) - \hat{r}_c(t, f_c; \omega_i, \Omega_j))^2 \tag{3.5}$$

where $\hat{r}_c$ is the denoised signal estimated from R_n using a linear transformation:

$$\hat{r}_c = HR_n \tag{3.6}$$

This problem is then reduced to the least squares estimation and the solution can be shown to satisfy the following equation (in matrix form):

$$(R_n R_n^T)H = R_n^T r_c \tag{3.7}$$

where $R_n R_n^T$ is the autocorrelation of the noisy speech and $R_n^T r_c$ is the cross-correlation of the clean and noisy speech. The auto and cross correlation matrices may also include lags larger than zero which can be helpful in such applications where reverberation and other temporal distortions need to be eliminated. For additive noise, however, we found that adding more lags does not improve the performance. Another significant finding is that the quality of the reconstructed speech improves significantly when extended to the space of nonlinear Volterra mappings. This is probably because of the non-stationary character of the speech signal, which necessitate the denoising filter be either a time-varying linear function or a nonlinear filter capable of capturing the dynamics of the underlying input signal. The former idea has been implemented before using Kalman filter [33], while the latter approach was attempted through the application of neural networks and Volterra series [24,34]. In this work, we used the second order Volterra representation of Noisy speech in the modulation domain as the input to all mappings denoted by R_n and computed in 3.5. Once the coefficients of the nonlinear Volterra filter are estimated, the filter is

applied on the original noisy speech, and then the output that is in spectro-temporal domain will be transformed into acoustic waveform.

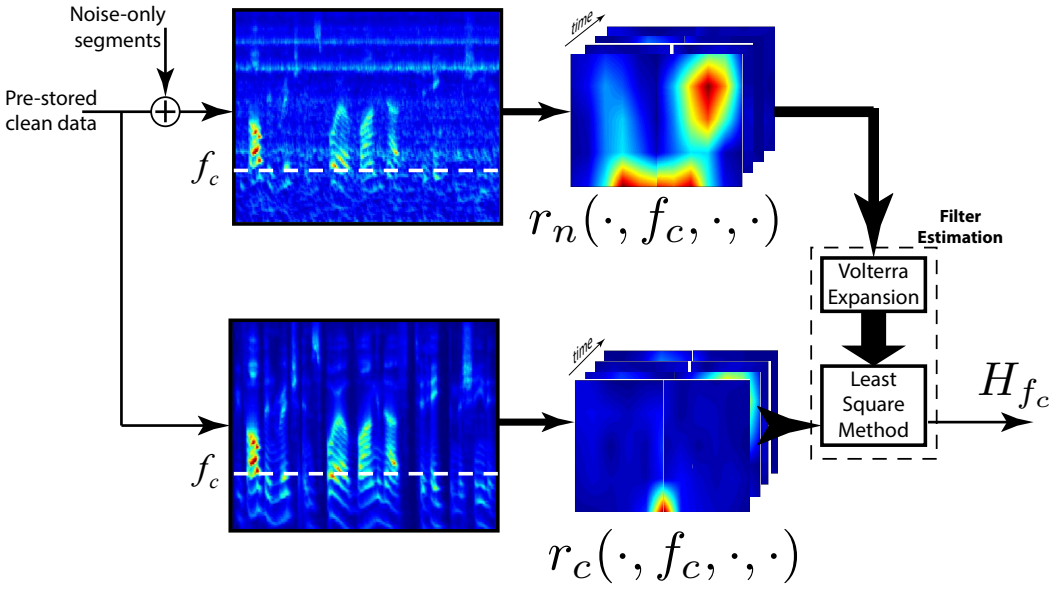

Figure 3.4: The nonlinear filter parameters are optimized by least squares method in such a way that the noisy representation of training data is mapped to the corresponding clean representation.

3.4 Evaluation Results

We evaluated the performance of the proposed nonlinear spectro-temporal modulation filtering (NSTMF) method using subjective tests and compared the quality of the denoised signal to the original noisy and also to denoised samples using the spectro-temporal modulation filtering (STMF) method suggested by Mesgarani *et al.* [26].

The results below were computed from noisy speech data, where the adaptive estimates of the nonlinear mapping were computed for three different types of noise extracted from Noisex [33] and averaged over ten clean speech samples from TIMIT [34]. The noise signals were: Pink, f16, Jet and Babble noise. The test material was prepared at four SNR ranges, -5, 0, +5 and +10 dB. We conducted subjective quality evaluation tests using mean opinion score (MOS) [35]. In the subjective quality tests, twenty subjects were asked to score the quality of the original and denoised speech samples between one (bad) and five (excellent).

Table 3.1 shows the average MOS results of ten subjects for three conditions: degraded speech, enhanced using the STMF [26], and the current NSTMF. The results are reported separately for different SNR and noise types. For the the stationary noises (the first three) the improvements in quality were almost significant, while for the nonstationary babble noise we only saw a slight improvement. This observation is consistent with the intrinsic assumption for the proposed algorithm that the statistical characteristics of the noise does not vary over time. The improvements for the proposed algorithm almost in all conditions beat that of the older STMF

method.

3.5 Discussion

We presented a speech enhancement method that took advantage of discriminatory capabilities of cortical representation of sounds. In contrast to standard STFT features typically used in enhancement systems the cortical features specially those corresponding to lower rates encode information from much longer windows in the input signal some times in the order of a second. This makes these features much richer in terms of the information they encode about spectro-temporal evolution of speech signals.

3.5.1 Size Issue

This intrinsic richness of cortical features comes at the cost of an overcompleteness manifested through the large size of the representation which in turn results in a computational burden when working with these features. A natural question that can be asked here is that if there exists other spectro-temporal representations of sound possibly that can describe speech and distinguish them from other types of signals but possibly more compact than the cortical basis. We will try to answer this question later in Chapter 5 through the use of non-negative matrix factorization framework.

Table 3.1: Mean Opinion Score on a scale of 1 to 5 averaged over 20 subjects for three conditions: (1) Original noisy speech (2) enhanced speech using STMF and NSTMF.

Noise type, SNR	Noisy	STMF	NSTMF
Pink, +10dB	3.6	3.9	4.3
Pink, +5dB	2.8	3.1	4.0
Pink, 0dB	2.0	2.1	2.5
Pink, -5dB	1.1	1.4	1.6
Jet, +10dB	3.7	4.0	4.3
Jet, +5dB	3.0	3.5	4.0
Jet, 0dB	2.0	2.3	2.4
Jet, -5dB	1.0	1.2	1.8
F16, +10dB	3.6	4.3	4.7
F16, +5dB	2.8	3.1	3.8
F16, 0dB	2.2	2.6	3.3
F16, -5dB	1.8	1.8	1.4
Babble, +10dB	3.5	3.7	3.9
Babble, +5dB	2.7	2.8	3.1
Babble, 0dB	2.0	2.0	2.4
Babble, -5dB	1.7	1.7	1.6

3.5.2 Adaptive vs Nonadaptive

The proposed method involved a stage of voice activity detection followed by a transformation in cortical domain applied on noisy segments. The input segments identified as non-speech were used to estimate the filters that could suppress the noise in cortical domain when applied on noisy signal. In our implementation, the filter estimation stage was performed in an offline fashion in the sense that assuming the noise statistics did not change significantly over time we collected all noise-only segments first and used them to compute the filter coefficients in a batch-mode using equation (3.7).

This introduce a lag equal to the duration of the noisy input to the enhancement scheme which might not be acceptable in some specific applications. Since the filters themselves are linear transforms one can think of computing them in an adaptive fashion. In this sense, once the cortical features for a segment in noisy signal are extracted the filter coefficients would be updated depending on wether the frame is identified as speech or nonspeech. The adaptive implementation would not only address noise types with slowly varying statistics over time but also seems to be biologically more plausible.

We also considered a certain form of nonlinearity for the feature transformation i.e. Volterra expansion followed by a linear transform. Studying other forms of nonlinearity would be of interest for the future direction of this work.

Chapter 4: Coherence-based mask estimation for speech enhancement

4.1 Overview

In the previous chapter, we saw how traditional speech enhancement methods can take advantage of our knowledge about feature extraction function of the auditory system through the use of a model suggested for sound representation in the primary auditory cortex. In this chapter we will put the feature extraction problem aside and focus on the feature modification stage. While traditional modification methods in the field mostly addressed this problem in a pure statistical framework in which the goal was to estimate clean speech representation from its sum with other independent processes (noise), during last decade, the attention of the scientific community has turned to the functionality of human auditory system as the biological means for speech perception. Since the notion of *audible distortion* was introduced in [36] and was taken into account to improve the intelligibility of speech, numerous other methods motivated by psychoacoustic studies have emerged in the field.

Within this mindset, another source of inspiration has been special capability of

humans and other animals auditory system in detecting, identifying and tracking sounds generated by a specific source in presence of sounds coming from other sources i.e. the auditory scene analysis (ASA). ASA has been defined as processes by which sequential and concurrent acoustic events are analyzed and organized into auditory streams. These streams are perceived by the listener as a coherent entity and, as such, can be selectively attended to among other sounds. Speech enhancement problem can be viewed as a special case of auditory scene analysis when the listener aims to attend to a speech signal in a noisy environment. This can create new opportunities for innovations in the enhancement field by adapting numerous hypotheses and models proposed in the ASA context.

While the biological processes in the brain underlying ASA determining which components and attributes in a mixture belongs to a certain source is yet to be fully understood, numerous hypotheses and models have tried to explain the neural basis of auditory perception in central auditory system and specially the auditory cortex based on neurophysiological data and psychophysical observations.

A prominent hypothesis in the field is the "population separation" theory of auditory streaming which suggests that sound elements segregate into separate streams whenever they activate well-separated populations of auditory neurons that are selective to frequency or any other sound attributes that have been shown to support stream segregation [37, 38]. One short-coming about this model is that it cannot account for the observed influence of the relative timing of sounds on streaming percepts. For example, the population-separation hypothesis predicts that both alternating and synchronous tones that differ widely in frequency should be heard as

separate streams. This prediction is contradicted by psychophysical and neurophysiological data [39] demonstrating that sequences of tones that are separated by an octave or more are still heard as a single stream if the tones are synchronous or, more precisely, fully coherent in time.

To address this shortcoming, another model was recently proposed by Shamma et al [40] that highlights the role of temporal coherence in auditory streaming and will be the center of our focus throughout this chapter. The two main arguments of this model are as following:

1. The formation of auditory streams depends fundamentally on the temporal coherence of responses of neural populations in the auditory cortex encoding various features of the sound. In particular, it is hypothesized that the temporal coherence between features is used as criterion to link those produced by the same sound source, while simultaneously separate them from others produced by other sources.

2. Attention plays role in the auditory scene analysis through enhancing responses to different sound features, and thus modifying the neural representation and ultimately the perceptual saliency of these features. In this way, the notion of feature-based attention is introduced and discussed that in situations where at least one distinctive feature of the target stream is sufficiently salient to be selectively attended to by the listener (called cue hereafter) it can serve as the anchor that points to and can be used to bind other features that are

temporally coherent with it.

A schematic of the this model illustrated in Figure 4.1 demonstrates how different attributes of the sound mixture such as spectrotemporal patterns, pitch tracks and location information are extracted in the feature and how their temporal dynamics is used trough a coherence analysis stage to group ensembles to form streams. It also shows how selective attention plays role on each of these stages.

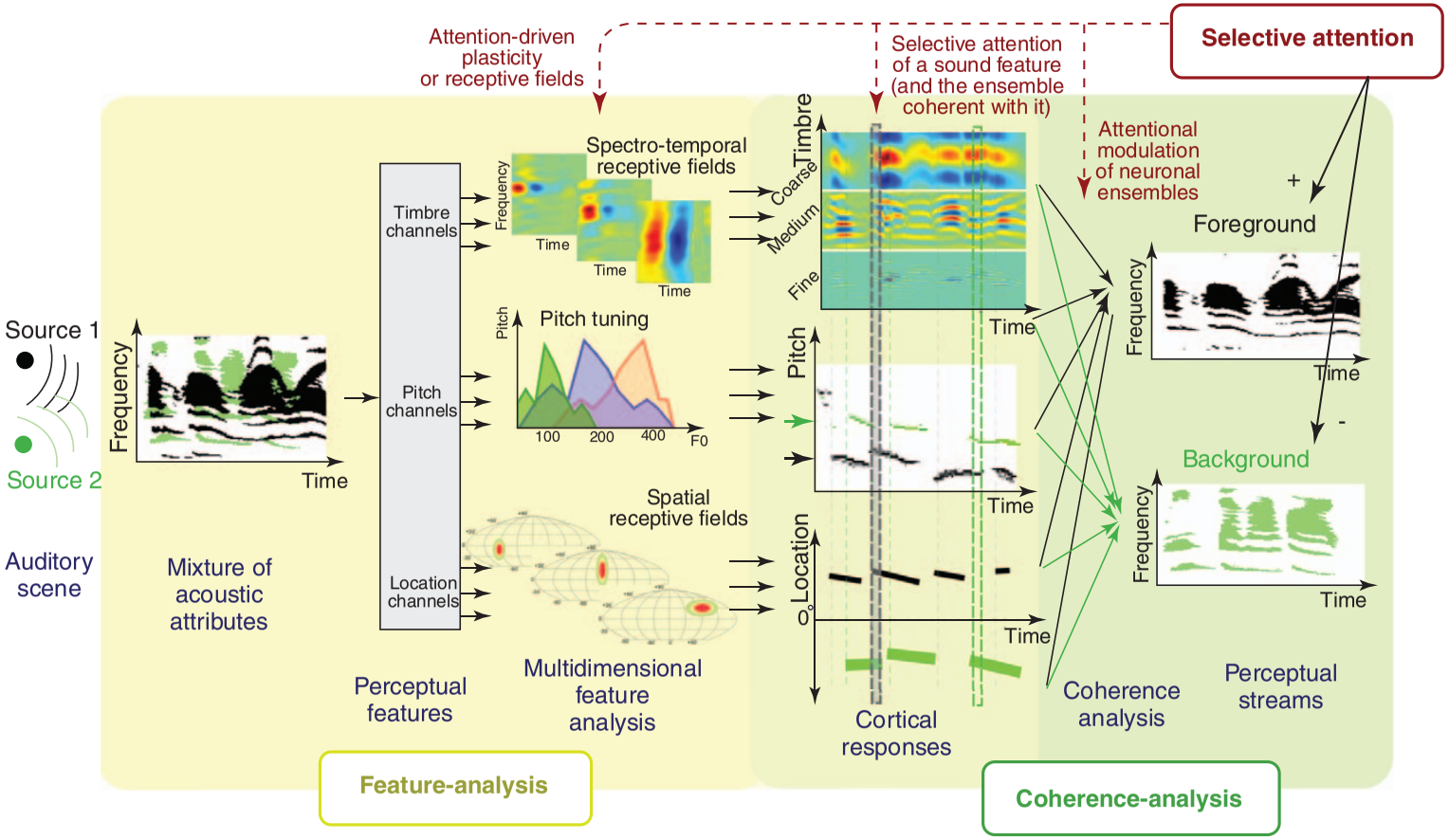

Figure 4.1: Schematic of the coherence-based model of auditory stream formation.

Inspired by this model of feature-based attention, we will suggest a mask-based enhancement method as presented in [41] that uses a measure of coherence between available dynamical cues and acoustic features to control the gain coefficients suppressing/enhancing those features. Previously, in works such as [42] and [43], cues in the form of visual sensory data (i.e. mouth shape) and those extracted from the acoustic waveform itself such as pitch tracks have been used in forming the gain functions. The distinctive feature of this work is that, the introduced model provides a multimodal framework in which single or multiple cues of different types can simultaneously or individually be fused into the enhancement system.

Figure 4.2 illustrates a schematic of the proposed method based on coherence analysis. The analysis stage decomposes the time-domain waveform into time-frequency features describing the noisy mixture in a higher dimensional space. The short-time coherence between the decomposed features and the cues are calculated. The coherence values are then translated into gain coefficients through nonlinear mappings. The gain functions are smoothed over time and finally the cleaned waveform is reconstructed using modified features in the synthesis stage. Throughout the following subsections, we will explain the mathematical formulation of auditory-inspired spectral weighting rule (AISWR) and elaborate the notion of temporal coherence. In the result section, we show two examples for which loudness and pitch tracks are used as the cue to enhance the perceptual quality of the speech in noise.

4.2 Auditory-Inspired Spectral Weighting Rule

In a single-acoustic-channel noisy environment, consider the time-frequency representation of a speech signal, $S(k,m)$ (with k and m respectively denoting time and frequency) that is corrupted by statistically independent background noise $N(k,m)$. The noisy mixture $X(k,m)$ can be represented as:

$$X(k,m) = S(k,m) + N(k,m) \tag{4.1}$$

The objective of a spectral weighting rule is to estimate the speech spectrum as follows:

$$\tilde{S}(k,m) = G(k,m)X(k,m) \tag{4.2}$$

A common form to construct Wiener-like filter $G(k,m)$ is as following:

$$G(k,m) = \xi_m(\varepsilon(k,m)) \tag{4.3}$$

where $\varepsilon(k,m)$ represents a measure of noise level at time k, and ξ_m a nonlinear mapping effective in suppressing background noise in the m-th subband but at the expense of speech distortion. Indeed, the measure $\varepsilon(k,m)$ must reflect the degree of speech signal dominance over noise in the m-th subband channel at time k, and its transform $G(k,m)$ should take values close to one when the likelihood of speech signal being completely dominant is high and vanish to zero at low SNR. Estimated subband SNR is a common example of such measure previously suggested in [44] in a single-microphone scenario. Alternatively, in a multimodal framework, where there is access to a dynamic cue feature of the target source, short-time coherence of

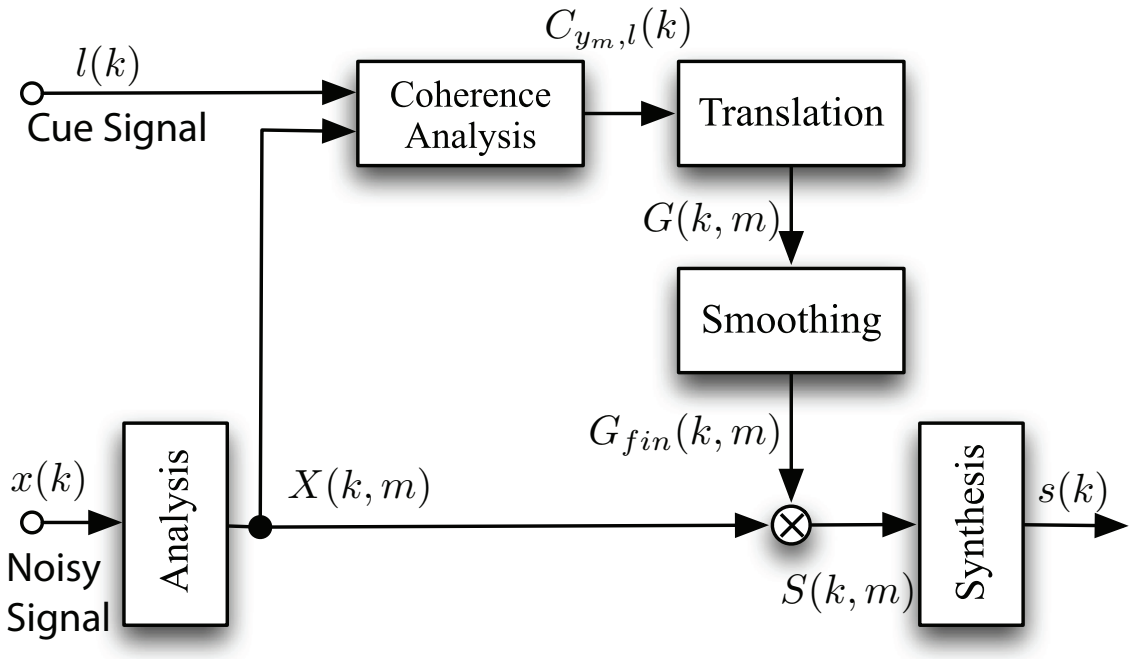

Figure 4.2: Schematic for AISWR

subband components with such a feature can serve as a measure of target superiority, i.e. $\varepsilon(k,m)$. In this way we will have:

$$G(k,m) = \phi_m\big(C_{X_m,l}(k)\big) \tag{4.4}$$

With $l(k)$ and $C_{X_m,l}(k)$ respectively denoting the cue signal and the short-time coherence between the cue signal and the m-th subband feature at time k, and $\phi_m(\cdot)$ a nonlinear function translating coherence values into appropriate gain coefficients. In the following two subsections, we will explain how to quantize the temporal coherence and compute the translation mappings.

4.2.1 Temporal Coherence

The term *temporal coherence* between two measured quantities is intuitively recognized as the extent their values evolve together over time. There exist several metrics suggested to quantize this type of association. The correlation coefficient of two variables can be a measure of coherence when referring to linear dependencies. Mutual Information (MI) is another measure that quantizes the broader case of statistical dependence between random variables. For two continuous random variables X and Y with joint and marginal densities $f_{X,Y}$, f_X and f_Y, Shannon defined MI defined as:

$$I(X,Y) = \iint f_{X,Y}(x,y) \ln \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy \tag{4.5}$$

MI has already been used to measure similarity in the context of clustering and feature selection in [45] on the basis that features belonging to the same objects must have strong statistical dependence. The necessity of tracking statistical dependence becomes specifically more evident in our application noting that the features in hand might be of different natures and hence their simultaneous changes over time might not be captured only from their second order statistics. An example of this phenomenon is the case when two signals with correlated envelopes (both following the same vocal tract shape changes) are modulated at different frequencies and their second order correlation is simply zero. For that, we use MI as a metric of coherence to extract information about acoustic features and to modify them.

4.2.2 Mutual Information Estimation

In applications, one usually has the data available in form of N sample points (x_i, y_i), $i = 1, \ldots, N$ which are assumed to be i.i.d. realizations of the underlying joint density $f_{X,Y}$. Since the underlying joint densities are unknown MI should be estimated from the available sample points. Among existing algorithms to estimate $I(X, Y)$, we adapted a k-nearest neighbor (KNN) estimator proposed by Kraskov *et al.* [46].

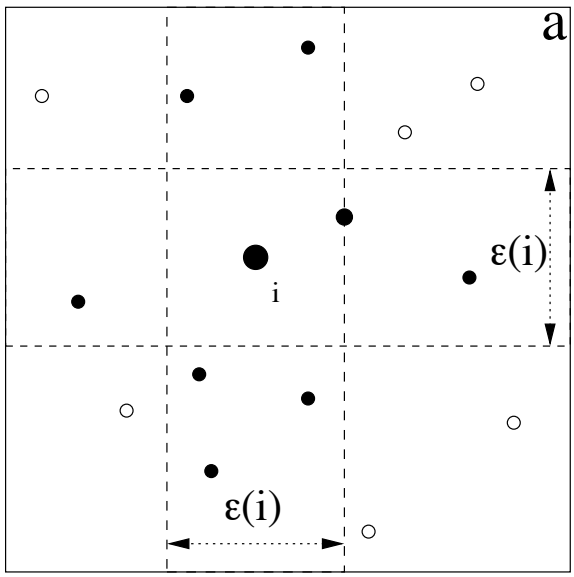

Figure 4.3: Determination of $\epsilon(i)$, $n_x(i)$ and $n_y(i)$ for $k = 1$. In this example, $n_x(i) = 5$ and $n_y(i) = 3$

With some arbitrary norm defined on the spaces spanned by X, Y and a maximum norm for $Z = (X, Y)$ i.e. $\|z - z'\| = \max\{\|x - x'\|, \|y - y'\|\}$, the method first ranks for each point $z_i = (x_i, y_i)$, its neighbors by distance $d_{i,j} = \|z_i - z_j\|$:

$d_{i,j_1} \leq d_{i,j_2} \leq d_{i,j_3} \leq \ldots$ with same ranking in the subspaces X and Y. By $\epsilon(i)/2$ being the distance from z_i to its k-th neighbor, and $\epsilon_x(i)/2$ and $\epsilon_y(i)/2$ the distances between the same points projected into X and Y subspaces, the method then counts the number $n_x(i)$ of points x_j whose distance from x_i is strictly less than $\epsilon(i)/2$ and similarly for y instead of x as shown in Figure 4.3. The mutual information is then estimated by:

$$I(X,Y) = \psi(k) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle + \psi(N) \tag{4.6}$$

With $\langle \cdots \rangle$ denoting averages both over all $i \in [1, \ldots, N]$ and over ll realizations of the random samples i.e.

$$\langle \cdots \rangle = N^{-1} \sum_{i=1}^{N} E[\ldots(i)] \tag{4.7}$$

and $\psi(x)$ being the digamma function, $\psi(x) = \Gamma(x)^{-1} d\Gamma(x)/dx$.

The KNN estimator for MI has been shown to be data-efficient and effective in capturing nonlinear dependence [47]. For dynamic signals such as speech for which the statistical characteristics change significantly over time, it is natural to compute the coherence between features over short time windows. In this way short-time coherence, $C_{x,y}(k)$ is defined as the MI estimated using the samples pairs of the two signals x and y in a window of appropriate length centered at time k.

4.2.3 Translation to Gain Coefficients

Once the short-time coherence values are calculated, they should be transformed to correct gain coefficients through the functions ϕ. Roughly speaking,

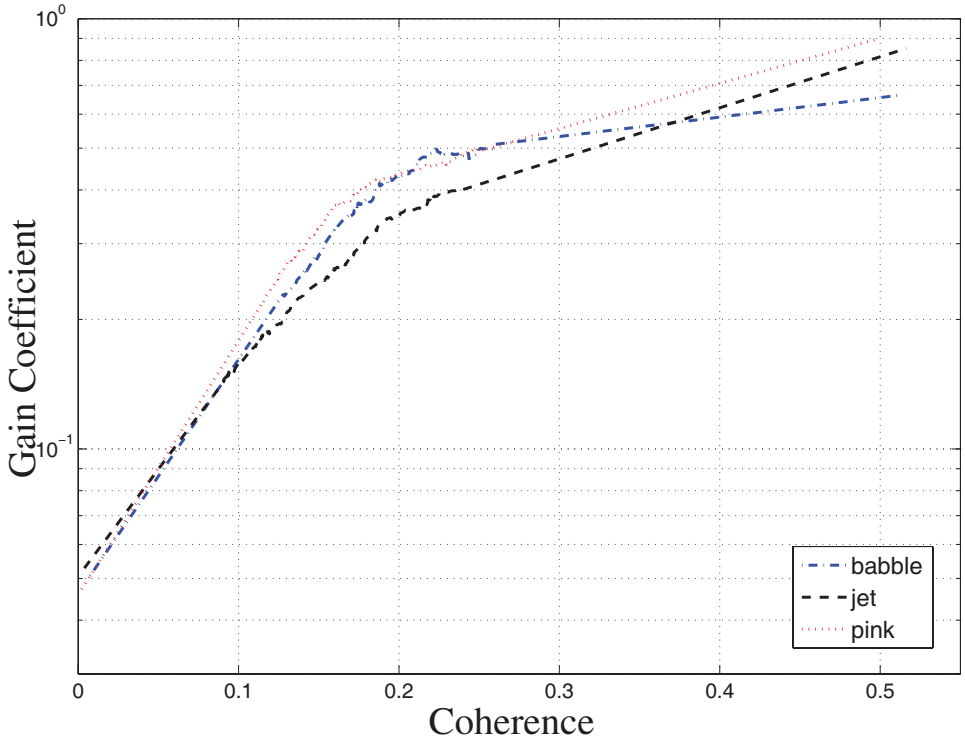

Figure 4.4: mappings ϕ_m computed for a specific feature using three different noise types, white, jet, and pink

the larger the value of the coherence is, the larger expected contribution of the speech signal in the mixture and consequently the value of the gain should be. In general, due to the unknown dependency of the cue to the target signal, unlike wiener-filtering approaches, it is hard to obtain closed form functions relating coherence values to gain coefficients. Because of that, to formulate these relations we follow a supervised learning approach.So given the cue signal(s) and the mixture decomposition, ϕ_m is trained offline on some noisy speech signals along their clean versions. For that, the ground truth gain coefficients are calculated, $G(k,m) = S(k,m)/X(k,m)$, a scatterplot is built with the pairs of gain coefficients and coherence values, $(G(k,m), C_{X_m,l}(k))$, and then smoothed, giving ϕ_m.

An important aspect of these mappings is that they are empirically found to follow same trends for various types of noise. This is in fact the key point that makes this technique independent of the noise characteristics. Figure 4.4 demonstrates such invariability by illustrating the mappings for a specific subband feature and the loudness signal (see section 4.3) computed for three different types of noise. Needless to mention, this invariability does not necessarily exist across different features as emphasized in (4.4) by the superscript m. In the final stage, the gains in each channel are smoothed in time by passing through a low-pass filter with a cutoff frequency matching the natural bandwidth of the corresponding feature. The logic behind this is the fact that rapid fluctuations in the gain has been shown to result in audible artifacts in the reconstructed waveforms as discussed in [48].

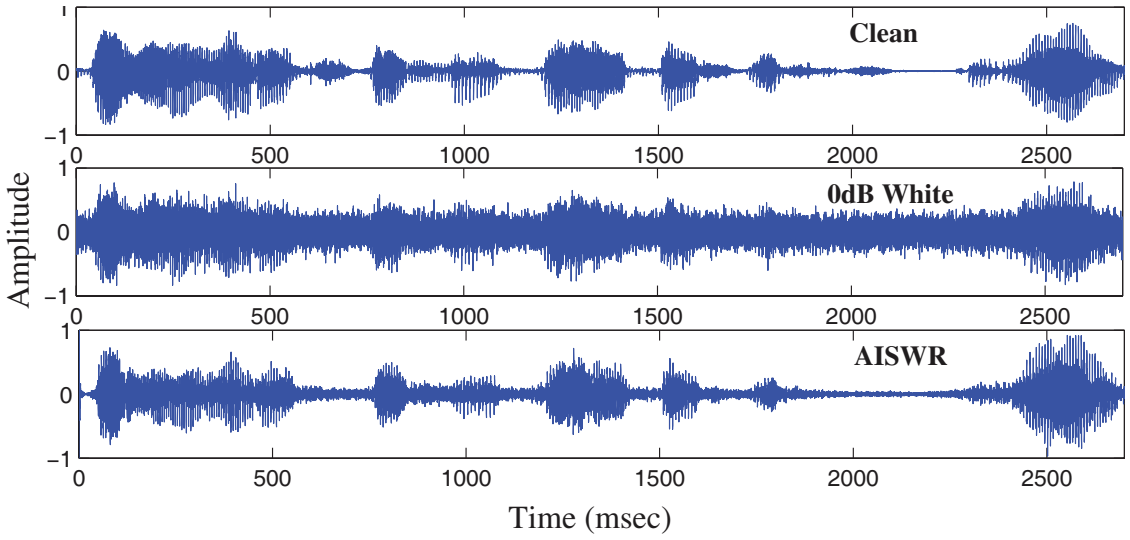

Figure 4.5: A clean speech sample chosen from TIMIT (above) noisy version in 0dB white noise (medium) corresponding signal enhanced by AISWR (bottom)

4.3 Results

In this section, we describe proof-of-concept examples for speech enhancement with the aid of two different types of cue signals. In both cases, objective tests were performed to assess the effectiveness of the proposed method. We compared AISWR against two other methods, the MMSE log-spectra amplitude estimation introduced in [44] and a recently proposed method called Block Thresholding (BT) based on adaptive wavelet denoising [49]. An objective metric, Enhanced Modified Bark Spectral Distortion (EMBSD) was used to evaluate the quality of the enhanced speech signals. This measure introduced in [50] is an improved version of BSD measure shown to have higher correlations with subjective Mean Opinion Score (MOS) of the perceptual quality of speech signals. Experiments were done using 20 speech sentences randomly selected from TIMIT corpus and corrupted with different types

of noise from NOISEX92 dataset [51] at four different SNR levels [0-15dB]. In all these experiments, we chose auditory spectrogram for the feature analysis stage. This frequency-time representation of the sound, inspired by the early stages of human auditory pathway, consists of almost constant-Q filters along the tonotopic (logarithmic frequency) axis followed by a nonlinear function capturing the envelope in each subband. The sound is decomposed into 128 real, positive-valued features evolving in time.

In the first example, the cue was chosen to be the power signal of the mixture representing the loudness of the target speech. We used three noise types (white, jet and babble) for which the power did not change significantly over time so that the cue was only correlated with the target source. In computation of STCC, 250 ms windows were chosen according to syllabic rate of speech. Figure 4.5 depicts a clean speech signal selected from TIMIT corpus along with the corrupted one in white noise (SNR=0dB), and the enhanced version using AISWR with loudness signal.

In the second example, we used for cue the pitch tracks extracted from the noisy speech signal. For that, first the values of the fundamental frequency, $f0$, were computed for each time frame in the mixture representation. To generate the cue signal at each time window, instead of using the pitch values themselves, we picked three subband channel outputs in the mixture representation corresponding to the first three harmonics of the salient calculated $f0$. In fact, the attended feature in this example was part of the representation itself that likely more correlated with the target speech signal. We used a publicly available package, praat [52] to analyze and extract $f0$ values from the noisy segments. The pitch estimation algorithm in

this package is pretty robust to a broad range of noise types. Thus the subband signals at harmonics with salient periodicity can serve as a reliable cues regardless of stationarity of the interfering signal. To show this, we chose two non-stationary noise types for this example, i.e. city and babble.

In the translation stage, separate mappings were trained for different values of $f0$ due to the fact each value mandates use of different channels as cue. This lead to 95 different mappings according to different pitch values in the range [50Hz-450Hz]. The mean EMBSD improvements for the three methods are reported in figure 4.6. AISWR outperformed the other two methods in all conditions when loudness was used as cue.

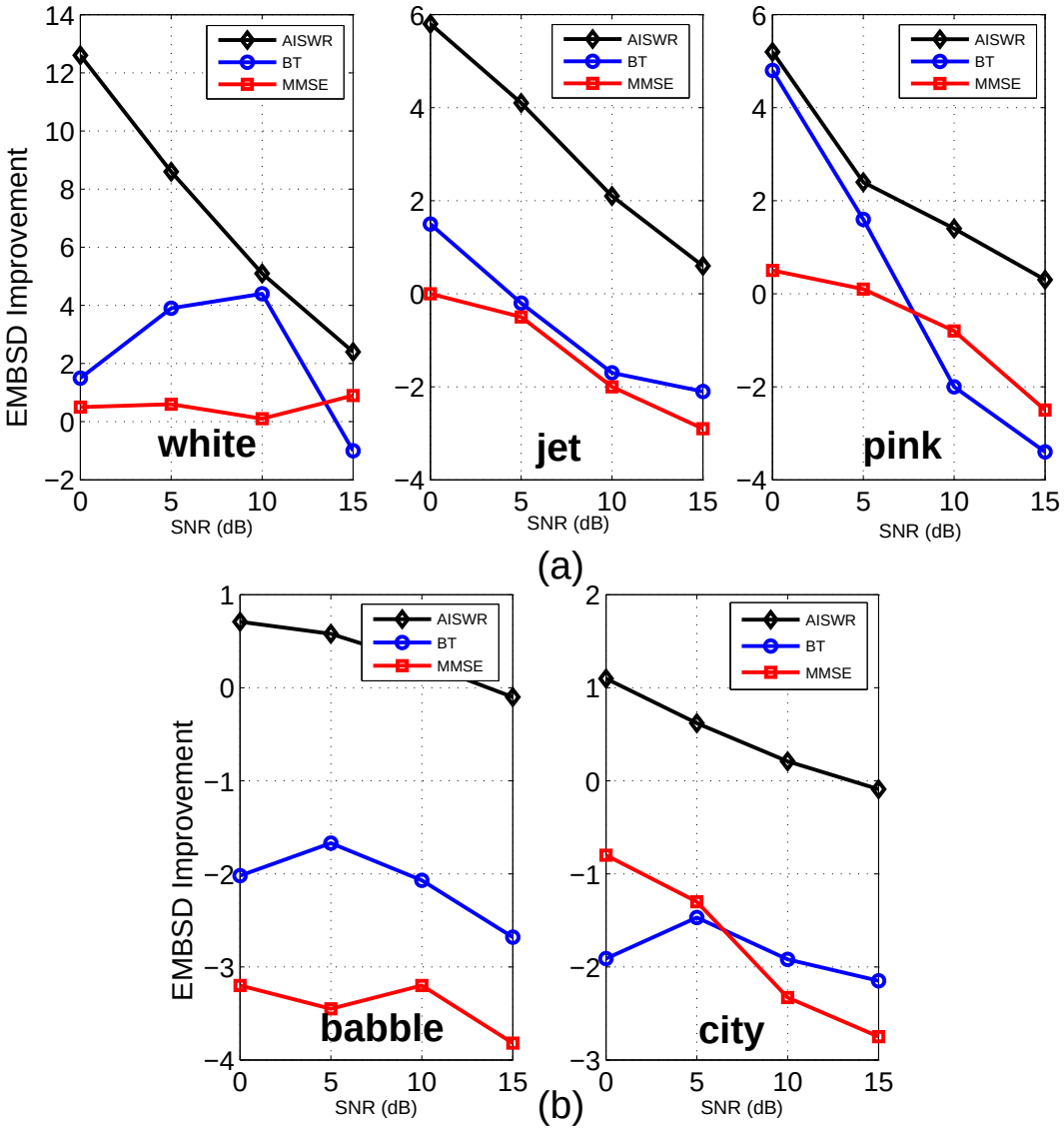

Figure 4.6: (a) loudness signal used as cue with three stationary noise types (b) pitch tracks used to extract the cue for two non stationary noise types.

4.4 Discussion

An auditory-inspired framework for speech enhancement was described in this chapter. This framework inspired by a model of attention that is hypothesized for

auditory scene analysis creates the capacity to capture information from cue signals and use it to modify acoustical features in the enhancement system. We used temporal coherence between the cue signals and the acoustical features to compute the gain coefficients in our spectral weighting enhancement scheme. Computation of coherence through a nonparametric estimate of mutual information between allows us to integrate one or multiple cue signals possibly in different modalities to the enhancement system in a unified framework. Two examples we used to show the effectiveness of the the proposed method were loudness signal and extracted pitch tracks. We showed through objective evaluations of sample audio files that the method can be effective in improving the speech quality.

We should here emphasize on the importance of the cue signal choice. It is easy to see that for a cue signal that is independent from the target source (speech), the method would result in meaningless gain functions as the coherence would always be measured as zero. So the more this cue signal carries information about underlying dynamic characteristics of the source the better it can be used as an anchor to detect target dominance in the acoustical features.

Another consideration about the proposed method is the choice of the estimator for measuring temporal coherence. As we saw the MI estimation algorithm involves finding k-th nearest neighbor which can be computationally expensive for certain applications. A future direction for this work would be considering other measures for statistical dependence that can be computed more efficiently or customized specifically for speech signals.

Chapter 5: Speech Enhancement Using Convolutive Nonnegative Matrix Factorization with Cosparsity Regularization

5.1 Overview

In the previous two chapters we introduced two enhancements schemes in which the analysis and the modification stages were designed by looking into models suggested for the auditory system. We saw how detecting certain spectro-temporal patterns that could differentiate speech and noise could serve the enhancement and also how temporal coherence between extracted features may be used as a criterion to separate noise from speech. A common aspect of these two schemes was that in both of them the analysis stage took place in cascade with the modification stage in a feed-forward manner. However we do not have any evidence suggesting that the auditory system is working on the same basis. As mentioned earlier there is no agreement among the neuroscience community about the feature extraction role of the auditory cortex [11] as some hypothesized about neurons being *descriptors* (as opposed to detectors) whose characteristics changes in response to the acoustic stimuli over time [53]. Indeed it has been shown that units even as low as hair cells receive feedback from higher regions in the auditory systems that modulates their

transfer functions [54].

An advantage of such an adaptive description scheme to represent the stimuli would be "compactness" which is of huge importance in any biological system with a limit in resources. Inspired by these aspects of the auditory system, in this chapter we consider an enhancement scheme in which:

1. spectro-temporal patterns are represented by descriptors that are optimized for speech signals and can adapt themselves for the specific given noisy samples

2. temporal correlations between identified features are used as a criterion to group them and separate noise from speech.

3. the above two processes take place in conjunction together mimicking their counterparts in the auditory system.

Based on the work presented in [55], we found nonnegative matrix factorization (NMF) framework plausible for this purpose. NMF-based enhancement techniques have recently been investigated and gained interest due to their capability in handling non-stationary noise types common in real-world applications. Standard NMF first introduced by Lee and Seung [56] simply aimed to approximate a non-negative matrix $X \in \mathbb{R}_{\geq 0}^{M \times L}$ as the product of two nonegative matrices $W \in \mathbb{R}_{\geq 0}^{M \times R}$ and $H \in \mathbb{R}_{\geq 0}^{R \times L}$ where $R \leq M$. With X being a magnitude spectrogram, NMF performs a linear basis decomposition storing the basis functions (atoms) in the columns of W and their corresponding temporal evolutions (activations) in the rows of H. Standard NMF ignores potential dependencies across successive columns of X. In order to account for temporal context convolutive nonnegative matrix factorization

(CNMF) was introduced by Smaragdis [57] by extending approximation of X as following:

$$X \approx \sum_{\tau=0}^{T-1} W(\tau) \overset{\tau\rightarrow}{H} \tag{5.1}$$

where $\{W(\tau)\}$ is a set of time-varying bases, $W(\tau) \in \mathbb{R}_{\geq 0}^{M \times R}$ and $H \in \mathbb{R}_{\geq 0}^{R \times L}$ contains the activities. The operator $\overset{\tau\rightarrow}{(.)}$ performs time-shifting by zero-padding of its operand with τ columns of zeros to the left and truncating that at the right to maintain correct dimensionality. Usually the approximation is done by solving a constrained optimization problem in which a divergence function between the input and its approximation is tried to be minimized subject to the nonnegativity of the congstructing matrices. To measure the reconstruction error, in this work we used the Frobenious norm, $\|.\|_F$, (i.e., the square root of summed squared matrix entries).

$$\underset{W(\tau),H}{\arg\min}\, D(X\|\hat{X}) \;\; \text{subject to} \;\; W(\tau), H \geq 0 \quad \forall \tau. \tag{5.2}$$

Now assuming additivity in the magnitude spectra domain, NMF-based speech enhancement methods usually aim to have each basis function (atom) in the final decomposition of the mixture only describe the speech or the noise spectrograms. In this way, enhancement would simply be achieved by combining speech components according to their corresponding activities in the mixture.

We should point out the fact that the additivity assumption about the magnitude spectra of the speech and noise, i.e. $X = S + N$, does not generally hold, however it has been shown that it is acceptable for the goal of source separation [57, 58].

Chickoki in [59] introduced a general NMF framework in which speech and noise separation in decomposition was achieved by means of regularizing the structure of the bases and their activations with regularization terms that penalize the reconstruction error in (5.2). Such regularizations usually take into account statistical characteristics, e.g. independence, or prior knowledge about the representation of the signal in hand (i.e. speech). A well-known example for the latter case is the sparsity of activations in CNMF representation of clean speech. It has been observed that preserving sparsity of speech activations usually results in better separation of speech and non-speech components [58] especially in presence of wideband noise.

A common measure to quantize sparsity is $\ell 1$-norm of speech activations over time. One issue about this and in general other sparsity measures used in this context is that they are only useful to minimize the global sparseness of the representation without accounting for how the occurrence of different components of clean speech are mutually correlated to each other. The importance of incorporating such information becomes specifically highlighted when the noise components resemble those of speech (e.g. second talker or babble noise) and hence are susceptible to activating speech bases. Wilson in [60] took advantage of prior information by assuming a normal distribution with known parameters for both speech and noise activations. In addition to being noise-dependent, one major shortcoming of this method was that their assumption implicitly mandated a certain value for both speech and noise signal powers. To incorporate prior information about activations without facing such issues, in this work, we introduce an extended notion of sparsity ,namely cosparsity. Having this measure quantize relative activation of bases pairs, the new regular-

ization on activations forces the components that are cosparse in clean speech not to co-occur in the denoised segment. We will show through some experiments how the new regularization can improve estimation of the speech spectrogram, $\hat{S}$, when used along with the standard sparsity term. In the following two sections, we first explain the cosparsity and the corresponding penalty function and then describe the regularized-CNMF method based on that.

5.2 Cosparsity

We define *cosparsity* between the activations of i-th and j-th components at the time instance l, in the following manner:

$$c_{ij}^{l} = \frac{h_{il}^2 + h_{jl}^2}{h_{il} h_{jl}} \tag{5.3}$$

with h_{il} being the activation of the i-th component at the time instance l. Note that for nonnegative activations the cosparsity measure always takes nonnegative values greater than or equal to 2. It takes its minimum value of 2 when activations are equal and approaches infinity as the ratio between the activations gets bigger and bigger or simply when the two components are cosparse. Note that the cosparsity is a symmetric relation and being only a function of relative strength, its value does not vary by scaling activations.

We shall maintain different levels of cosparsity among all pairs of components according to their record of cosparsity learned from an available clean speech dataset. This is done by prioritizing having larger cosparsity between pairs that are seldomly active at the same time in clean speech. We assume that there is a speech cor-

pus available for training that can be used to learn the prior knowledge on speech components. A basis, $\widetilde{W}_S(\tau)$, and the corresponding activation matrix, $\widetilde{H}_S(\tau)$ are pretrained on the speech corpora using standard CNMF.

We use the codebook activations, $\widetilde{H}_S$, to calculate a matrix P, keeping track of cosparsity of component pairs. In order to maintain the high cosparsity for pairs with corresponding low entries in P, we minimize a regularization term. This new term is basically the sum of the bounded inverse of cosparsity measure, $\frac{h_{il}h_{jl}}{h_{il}^2+h_{jl}^2}$, for all pairs weighted by the entries in P. Entries in P are between 0 and 1, where a value close to 1 reflects a cosparse pair and one close to 0 occurs when the activations are very similar. Having these properties in mind, we formed the entries in P as:

$$p_{ij} = (1 - \frac{\tilde{\mathbf{h}}_i.\tilde{\mathbf{h}}_j}{|\tilde{\mathbf{h}}_i||\tilde{\mathbf{h}}_j|})^{\zeta} \tag{5.4}$$

with $\tilde{\mathbf{h}}_i$ being the i-th row in the matrix $\widetilde{H}_S$ and $\tilde{\mathbf{h}}_i.\tilde{\mathbf{h}}_j$ and $|\tilde{\mathbf{h}}_i|$ respectively being the inner product and the $\ell 2$-norm of the vectors. The constant ζ simply controls the distribution of the entries of P on the interval $[0, 1]$. High values of ζ enforces cosparsity on smaller number of pairs with very high records of cosparsity while a very low value enforces cosparsity to all the pairs evenly.

5.3 CNMF with cosparsity regularization

Given a noisy speech spectrogram $X \in \mathbb{R}_{\geq 0}^{M \times n}$, the proposed regularized CNMF forms the estimate spectrogram as following:

$$\widehat{X} = \sum_{\tau=0}^{T-1} \begin{bmatrix} W_S(\tau) W_N(\tau) \end{bmatrix} \begin{bmatrix} \overset{\tau\rightarrow}{H}_S \\ \overset{\tau\rightarrow}{H}_N \end{bmatrix} \tag{5.5}$$

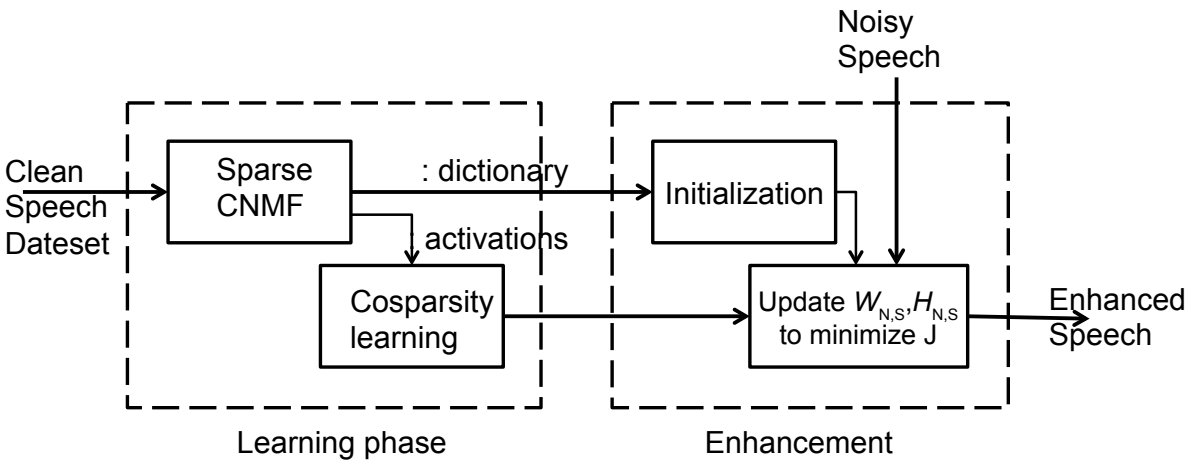

Figure 5.1: A schematic of CNMF with cosparsity regularization method

Here, $W_S(\tau) \in \mathbb{R}_{\geq 0}^{M \times R_S}$ and $W_N(\tau) \in \mathbb{R}_{\geq 0}^{M \times R_N}$ respectively denote speech and noise bases while $H_S \in \mathbb{R}_{\geq 0}^{R_S \times L}$ and $H_N \in \mathbb{R}_{\geq 0}^{R_N \times L}$ represent their activations. The estimate spectrogram $\widehat{X}$ is then computed by minimizing the following cost function with respect to the nonnegative basis and activation matrices:

$$\mathcal{J} \coloneqq \frac{1}{2}\|X - \widehat{X}\|_F + \alpha \cdot J_S(H_S) + \beta \cdot J_C(H_S) \tag{5.6}$$

where $J_S(H_S)$ and $J_C(H_S)$ respectively represent sparsity and cosparsity regularization terms computed as:

$$J_C(H_S) = \sum_{i \neq j} p_{ij} \sum_{l=1}^{L} {c_{ij}^l}^{-1} \tag{5.7}$$

and

$$J_S(H_S) = \sum_{l=1}^{L} \|\mathbf{h}_S^l\|_1 \tag{5.8}$$

with $\|\mathbf{h}_S^l\|_1$ denoting the $\ell 1$-norm of the l-th column in H_S and P being the cosparsity penalty matrix. α and β are two constants determining the amount of punish for the sparsity and cosparsity terms. A higher value for the constant α yields in lower value/number of active components for the speech part, and for the constant β results in a generally higher degree of cosparsity maintained between components. Similar to standard CNMF, the optimization will be performed through initializing the entries in each of these matrices and then a series of alternating updates on the basis and activation matrices according to multiplicative rules [61]. In order to preserve the nonnegativity of these matrices, this procedure updates them by gains that are a function of the terms in the corresponding gradient of the cost function $\mathcal{J}$. For any of these matrices say A, considering that the partial derivative matrix of the objective function with respect to elements in A can be decomposed into two nonnegative parts as:

$$\left.\frac{\partial \mathcal{J}}{\partial A'}\right|_{A'=A_{\text{old}}} = \nabla^+ - \nabla^- \tag{5.9}$$

In this way, we will have the multiplicative update rule as:

$$A_{\text{new}} = A_{\text{old}} \odot \frac{\nabla^-}{\nabla^+} \tag{5.10}$$

where $\odot$ is the Hadamard product (element-wise multiplication), and division between the matrices is also an element-wise operation. For the cosparsity term, these two nonegative parts ∇_C^+ and ∇_C^- can be element-wise derived as:

$$\frac{\partial J_C}{\partial h_{il}} = \sum_{j;i\neq j} \frac{p_{ij} h_{jl}^3}{(h_{jl}^2 + h_{il}^2)^2} - \sum_{j;i\neq j} \frac{p_{ij} h_{jl} h_{il}^2}{(h_{jl}^2 + h_{il}^2)^2} = \delta_{il}^+ - \delta_{il}^- \tag{5.11}$$

Thus following [59], the new multiplicative update for speech activations would be

expressed by:

$$H_S \leftarrow \left\langle H_S \odot \frac{W_S^T(\tau)\overleftarrow{X}^{\tau} + \nabla_C^-}{W_S^T(\tau)\overleftarrow{\widehat{X}}^{\tau} + \beta \cdot \mathbf{1}_{R_S \times L} + \nabla_C^+} \right\rangle_{\tau} \tag{5.12}$$

Since noise activations only appear in the error term of $\mathcal{J}$ the multiplicative update for them would be:

$$H_N \leftarrow \left\langle H_S \odot \frac{W_S^T(\tau)\overleftarrow{X}^{\tau}}{W_S^T(\tau)\overleftarrow{\widehat{X}}^{\tau}} \right\rangle_{\tau} \tag{5.13}$$

Following the same procedure, the basis matrix $W(\tau) = \Big[W_S(\tau)W_N(\tau)\Big]$ is also updated in the following way:

$$W(\tau) \leftarrow W(\tau) \odot \frac{X\overrightarrow{H}^{\tau T}}{\widetilde{X}\overrightarrow{H}^{\tau T}} \tag{5.14}$$

with $H = \begin{bmatrix} H_S \\ \\ H_N \end{bmatrix}$. We also normalize the basis columns after each multiplicative update so that they all have their $\ell 1$-norms equal to one. The entries in the basis are initialized by the pretrained codebook, $\widetilde{W_S}$. It should be noted that the initialization is necessary not only for its known importance in optimization but also as kind of a labeling of speech components whose pairs are supposed to have uneven degrees of cosparsity. All other three matrices are initialized with random nonnegative values. The alternation between updates on the basis and activations is continued until either the relative change in $\mathcal{J}$ is lower than 1% or the number of iterations exceeds 150 (whichever happens first). The speech spectra is then simply estimated as:

$$\widehat{S} = \sum_{\tau=0}^{T-1} W_S(\tau)\ \overrightarrow{H_S}^{\tau} \tag{5.15}$$

Using the phase info from the noisy spectra and the overlap-add method, $\angle X$, we then generate the enhanced waveforms.

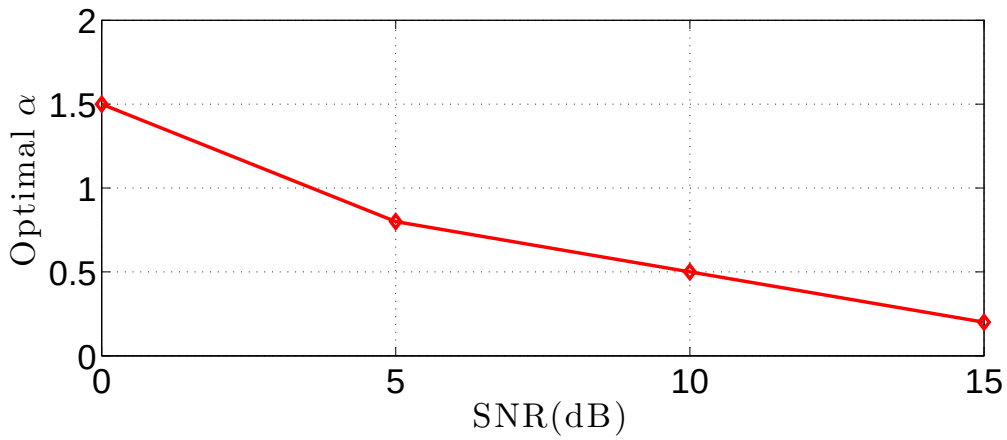

Figure 5.2: Optimal value for the sparsity term weight α computed for pink noise at different SNR levels.

5.4 Results and Experiments

In order to assess the performance of our proposed method, we used a speech corpora consisting utterances from the TIMIT database. The speech waveforms all sampled at their original rate, 16 kHz, were transformed to magnitude spectrogram (short-time Fourier transform) using 32 ms Hamming-weighted windows overlapped by 50%. The speech codebook and the cosparsity penalty coefficients were computed on about three minute of clean speech from randomly-chosen male and female speakers in TIMIT `train` subset. For the testing, we used 20 sentences from 10 male and 10 female speakers from TIMIT `test` subset. These sentences were corrupted by additive noise at four different SNR levels ranging from 0 dB to 15 dB. Three different noise types chosen from NOISEX dataset were evaluated in

our experiments, *Pink* (a stationary noise with energy uniformly spread over log frequency scale), and *City* and *Babble* as two cases of nonstationary noise.

Speech qualities were measured by a standard objective metric, perceptual evaluation of speech quality (PESQ) [62]. The measure was particularly developed to model subjective tests commonly used in telecommunications. However it has been commonly used in assessing quality of speech enhancement algorithms as well. PESQ takes values between .5 (bad) and 4.5 (no distortion).

In the experiments, we set $R_S = 100$, $R_N = 50$, and $T = 3$. In order to investigate how the quality of enhancement is effected by selection of the parameters α, β and ζ, the algorithm was run on corrupted samples in pink noise at different SNR levels using different combinations of these parameters. We considered $\alpha \in [0, 5]$, $\beta \in [0, 10]$ and $\zeta \in [0.1, 100]$.

For the sparsity term constant, α, the optimal value was computed by averaging the PESQ scores across speakers and different values of the two other parameters. For each SNR level, the value giving rise to the maximum quality was chosen as the optimal one. Figure 5.2 demonstrates these optimal values of α calculated for four different SNR levels. This result is consistent with the one reported in [58], and confirms that in lower SNRs, a more sparse reconstruction of speech results in a better quality.

Having set α to its optimal values for each SNR condition and averaging PESQ scores over different values of β, we observed that a value roughly equal to 20 for the parameter ζ almost always resulted in the highest scores. Finally, using these optimal values found for α and ζ, we looked at the average PESQ for different values

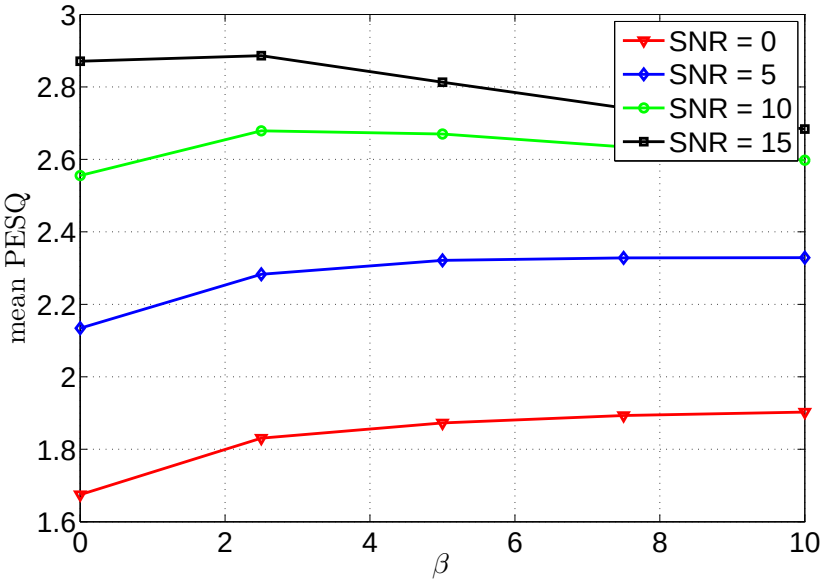

Figure 5.3: mean PESQ vs. cosparsity term weight, β.

of β. Figure 5.3 shows how the selection of β affects quality of reconstructed speech segments. It is observed that for all SNR levels except 15dB the maximum improvement is achieved for a value of β greater than zero suggesting the effectiveness of having cosparsity regularization term in the CNMF framework. Similar to the trend for the parameter α, it is observed here that as the intensity of noise is increased, the optimal value of beta also increase. This means that enforcing cosparsity relations between speech components becomes more essential as the noise gets stronger and able to activate speech components more frequently.

We finally compared our proposed method against the regular sparse CNMF and a baseline speech enhancement method. For the baseline method we chose

a spectral subtraction algorithm introduced in [63]. The comparison we did was best versus best, i.e. for the regular sparse CNMF, we simply set β to zero, and reported the highest mean PESQ score over all values of the parameter α while for ours it was the highest score over all values of α and all non-zero values of the parameter β ($\zeta = 20$). The results of the comparison are illustrated in figure 5.4. The improvement with respect to regular sparse CNMF is obvious for all three noise types especially in lower SNRs. Our method outperforms the baseline one for *Pink* and *City* noise. However, the results for *Babble* noise are somehow weaker. We believe this is related to speech-like statistical properties of this type of noise which poses a challenge to methods based on *a priori* knowledge of speech.

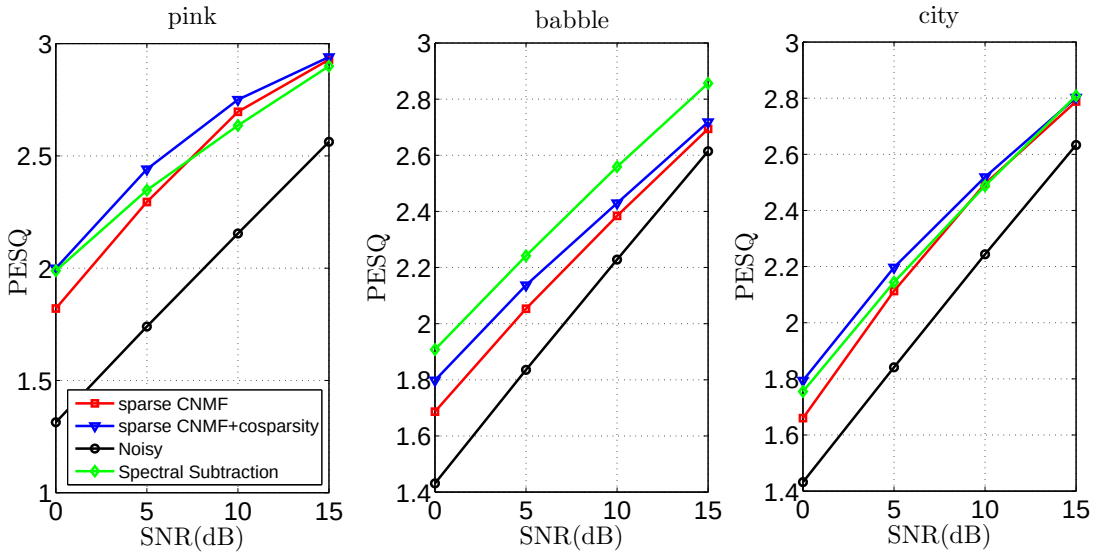

Figure 5.4: PESQ Improvements for different noise types.

5.5 Discussion

In this chapter, we investigated a offline speech enhancement scheme in which the spectro-temporal feature detection and modification blocks were merged into one step performed through the use of nonnegative matrix factorization of noisy speech representation. For this purpose, we took advantage of NMF framework to separate noise patterns from speech ones using a new regularization accounting for pairwise temporal coherence of the patterns, referred to as *cosparsity*.

In the learning phase, a dictionary of spectro-temporal patterns (features) was first generated from a general clean speech dataset. Then cosparse feature pairs were detected through a measure of pairwise coherences between them. In the enhancement phase, we enforced cosparsity between speech feature pairs using the information captured in the learning phase.

We discussed how selection of parameters can impact the quality of the reconstructed speech signals, and showed through objective evaluations that accounting for temporal coherence between features i.e. the new regularization can effectively improve quality of enhancement especially in low SNR conditions.

An advantage of this scheme is that the quality of separation can be easily improved when prior knowledge about the noise type or target speaker is available. This can be done through customized dictionaries for both speaker and noise parts. Similar to other offline NMF-based enhancement algorithms, a natural drawback of this method is that it's time and memory expensive making it unsuitable for real-time application. One future direction for this work would be adapting the existing online

NMF methods such as [64] possibly with a modified cosparsity regularization term exclusively for enhancement framework.

It should also be noted that the cosparsity regularization technique introduced in this chapter can be beneficial for any application dealing with part-based decomposition of input signals. A few examples are face recognition, text mining and bioinformatics.

Chapter 5: Conclusions

5.1 Thesis overview

Throughout this thesis we attempted to bring our knowledge about the auditory system to the field of speech enhancement in the hope of bridging the gaps between how humans handle the noise in auditory tasks and the engineering models that can be used in enhancement systems. We investigated different enhancement schemes that adapted auditory models in different stages of enhancement procedure. To show the importance of feature extraction in speech enhancement we presented a method that took advantage of sound representation in auditory cortex to isolate noise-only segments in noisy sentences and use them to construct the enhancement filters. We observed using spectro-temporal features along with simple linear modification rules could be effective for speech enhancement.

We then explored coherence-based model of attention in auditory scene analysis and saw how a measure of coherence based on mutual information between acoustical features and the cue signals could be used to form gain functions in a mask-based enhancement scheme.

We finally saw how the analysis and modification stages could be done simultaneously so that the detected spectro-temporal patterns are optimized for speech signals

and their temporal activities are taken into account to separate noise from speech.

5.2 Future directions

With the advances in our understanding about the auditory system we expect these methods to be revised and become more convergent with the new findings. We also believe devising novel engineering tools would be crucial to improve existing schemes in this context. For example in view of the recent progresses of deep learning methods in sound processing applications one can ask how to design new architectures that can serve as computational models for auditory scene analysis.

Bibliography

[1] K Uwe Simmer, Joerg Bitzer, and Claude Marro. Post-filtering techniques. In *Microphone Arrays*, pages 39–60. Springer, 2001.

[2] Martin Cooke, John R Hershey, and Steven J Rennie. Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1):1–15, 2010.

[3] Tuomas Virtanen. Speech recognition using factorial hidden markov models for separation in the feature space. In *INTERSPEECH*. Citeseer, 2006.

[4] Ji Ming, Timothy J Hazen, and James R Glass. Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation. *Computer Speech & Language*, 24(1):67–76, 2010.

[5] Jon Barker, Ning Ma, André Coy, and Martin Cooke. Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. *Computer Speech & Language*, 24(1):94–111, 2010.

[6] Yang Shao, Soundararajan Srinivasan, Zhaozhang Jin, and DeLiang Wang. A computational auditory scene analysis system for speech segregation and robust speech recognition. *Computer Speech & Language*, 24(1):77–93, 2010.

[7] Eric R Kandel, James H Schwartz, Thomas M Jessell, et al. *Principles of neural science*, volume 4. McGraw-Hill New York, 2000.

[8] Richard Lyon and Shihab Shamma. Auditory representations of timbre and pitch. In *Auditory computation*, pages 221–270. Springer, 1996.

[9] Xiaowei Yang, Kuansan Wang, and Shihab A Shamma. Auditory representations of acoustic signals. *Information Theory, IEEE Transactions on*, 38(2):824–839, 1992.

[10] Kuansan Wang and Shihab Shamma. Self-normalization and noise-robustness in early auditory representations. *Speech and Audio Processing, IEEE Transactions on*, 2(3):421–435, 1994.

[11] Israel Nelken, Alon Fishbach, Liora Las, Nachum Ulanovsky, and Dina Farkas. Primary auditory cortex of cats: feature detection or something else? *Biological cybernetics*, 89(5):397–406, 2003.

[12] T. Chi, P. Ru, and S.A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118:887, 2005.

[13] Didier A Depireux, Jonathan Z Simon, David J Klein, Shihab A Shamma, et al. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of neurophysiology*, 85(3):1220–1234, 2001.

[14] Nina Kowalski, Didier A Depireux, and Shihab A Shamma. Analysis of dynamic spectra in ferret primary auditory cortex. 1. characteristics of single unit responses to moving ripple spectra. Technical report, DTIC Document, 1995.

[15] Nina Kowalski, Didier A Depireux, and Shihab A Shamma. Analysis of dynamic spectra in ferret primary auditory cortex: Ii. prediction of, unit responses to arbitrary dynamic spectra. 1995.

[16] Taishih Chi, Yujie Gao, Matthew C Guyton, Powen Ru, and Shihab Shamma. Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America*, 106(5):2719–2732, 1999.

[17] Kuansan Wang and Shihab A Shamma. Spectral shape analysis in the central auditory system. *Speech and Audio Processing, IEEE Transactions on*, 3(5):382–395, 1995.

[18] Jae S Lim and Alan V Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

[19] Yariv Ephraim and Harry L Van Trees. A signal subspace approach for speech enhancement. *Speech and Audio Processing, IEEE Transactions on*, 3(4):251–266, 1995.

[20] Yariv Ephraim and David Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(2):443–445, 1985.

[21] Rainer Martin. Statistical methods for the enhancement of noisy speech. In *Speech Enhancement*, pages 43–65. Springer, 2005.

[22] Shihab Shamma. Encoding sound timbre in the auditory system. *IETE Journal of research*, 49(2/3):145–156, 2003.

[23] Mounya Elhilali, Taishih Chi, and Shihab A Shamma. A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech communication*, 41(2):331–348, 2003.

[24] G.B. Ginnakis and A.V. Dandawatw. Linear and non-linear adaptive noise cancellers. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '90.*, pages 1373–1376, 1990.

[25] Robert P Carlyon and Shihab Shamma. An account of monaural phase sensitivity. *The Journal of the Acoustical Society of America*, 114(1):333–348, 2003.

[26] Nima Mesgarani and Shihab Shamma. Denoising in the domain of spectrotemporal modulations. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(3):3, 2007.

[27] Majid Mirbagheri, Nima Mesgarani, and Shihab Shamma. Nonlinear filtering of spectrotemporal modulations in speech enhancement. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5478–5481. IEEE, 2010.

[28] Amir Hussain, Mohamed Chetouani, Stefano Squartini, Alessandro Bastari, and Francesco Piazza. Nonlinear speech enhancement: an overview. In *Progress in nonlinear speech processing*, pages 217–248. Springer, 2007.

[29] Nima Mesgarani, Malcolm Slaney, and Shihab A Shamma. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(3):920–930, 2006.

[30] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[31] Vladimir Vapnik. *The nature of statistical learning theory*. springer, 2000.

[32] Thorsten Joachims. Making large scale svm learning practical. *HT014602036*, 1999.

[33] K.K. Paliwal and A. Basu. A speech enhancement method based on kalman filtering. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, volume 12, pages 177–180, Apr 1987.

[34] Inhyok Cha and Saleem A Kassam. Interference cancellation using radial basis function networks. *Signal Processing*, 47(3):247–268, 1995.

[35] International Telecommunication Union. Telecommunication Standardization Sector. *Methods for Subjective Determination of Transmission Quality*. International Telecommunication Union, 1996.

[36] D.E. Tsoukalas, J.N. Mourjopoulos, and G. Kokkinakis. Speech enhancement based on audible noise suppression. *Speech and Audio Processing, IEEE Transactions on*, 5(6):497 –514, 1997.

[37] Susan L McCabe and Michael J Denham. A model of auditory streaming. *The Journal of the Acoustical Society of America*, 101(3):1611–1621, 1997.

[38] Yonatan I Fishman, Joseph C Arezzo, and Mitchell Steinschneider. Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *The Journal of the Acoustical Society of America*, 116(3):1656–1670, 2004.

[39] Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J Oxenham, and Shihab A Shamma. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, 61(2):317–329, 2009.

[40] Shihab A. Shamma, Mounya Elhilali, and Christophe Micheyl. Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 34(3):114 – 123, 2011.

[41] Majid Mirbagheri, Sahar Akram, and Shihab A Shamma. An auditory inspired multimodal framework for speech enhancement. In *INTERSPEECH*, 2012.

[42] I. Almajai and B. Milner. Visually derived wiener filters for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(6):1642 –1651, aug. 2011.

[43] S. Vishnubhotla and C.Y. Espy-Wilson. An algorithm for speech segregation of co-channel speech. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 109 –112, april 2009.

[44] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(2):443 – 445, apr 1985.

[45] A. Kraskov, H. Stögbauer, R.G. Andrzejak, and P. Grassberger. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70:278, 2005.

[46] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[47] Shiraj Khan, Sharba Bandyopadhyay, Auroop R Ganguly, Sunil Saigal, David J Erickson III, Vladimir Protopopescu, and George Ostrouchov. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2):026209, 2007.

[48] D.V. Anderson. A modulation view of audio processing for reducing audible artifacts. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5474–5477, 2010.

[49] Guoshen Yu, S. Mallat, and E. Bacry. Audio denoising by time-frequency block thresholding. *Signal Processing, IEEE Transactions on*, 56(5):1830 –1839, may 2008.

[50] Wonho Yang. *Enhanced modified bark spectral distortion (embsd): an objective speech quality measure based on audible distortion and cognition model*. PhD thesis, Temple University, Philadelphia, PA, USA, 1999.

[51] Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.

[52] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 5.3.11) [computer program], 2012.

[53] GHe al Recanzone, CE Schreiner, and Michael M Merzenich. Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *The Journal of Neuroscience*, 13(1):87–103, 1993.

[54] Florin Vranceanu, Guy A Perkins, Masako Terada, Robstein L Chidavaenzi, Mark H Ellisman, and Anna Lysakowski. Striated organelle, a cytoskeletal structure positioned to modulate hair-cell transduction. *Proceedings of the National Academy of Sciences*, 109(12):4473–4478, 2012.

[55] Majid Mirbagheri, Yanbo Xu, Sahar Akram, and Shihab A Shamma. Speech enhancement using convolutive nonnegative matrix factorization with cosparsity regularization. In *INTERSPEECH*, pages 456–459, 2013.

[56] D Seung and L Lee. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13:556–562, 2001.

[57] Paris Smaragdis. Convolutive speech bases and their application to supervised speech separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):1–12, 2007.

[58] Ruairí De Fréin and Scott T Rickard. Learning speech features in the presence of noise: Sparse convolutive robust non-negative matrix factorization. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–6. IEEE, 2009.

[59] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE, 2006.

[60] Kevin W Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4029–4032. IEEE, 2008.

[61] Wenwu Wang, Andrzej Cichocki, and Jonathon A Chambers. A multiplicative algorithm for convolutive non-negative matrix factorization based on squared euclidean distance. *Signal Processing, IEEE Transactions on*, 57(7):2858–2864, 2009.

[62] TP ITU and P Recommendation. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Recommendation ITU-T*, 2001.

[63] Yang Lu and Philipos C Loizou. A geometric approach to spectral subtraction. *Speech communication*, 50(6):453–466, 2008.

[64] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.