# ROBUST SIGNAL PROCESSING METHODS FOR MINIATURE ACOUSTIC SENSING, SEPARATION, AND RECOGNITION

By

Amin Fazel

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Electrical Engineering

2012

## ABSTRACT

## ROBUST SIGNAL PROCESSING METHODS FOR MINIATURE ACOUSTIC SENSING, SEPARATION, AND RECOGNITION

**By**

**Amin Fazel**

One of several emerging areas where micro-scale integration promises significant breakthroughs is in the field of acoustic sensing. However, separation, localization, and recognition of acoustic sources using micro-scale microphone arrays poses a significant challenge due to fundamental limitations imposed by the physics of sound propagation. The smaller the distance between the recording elements, the more difficult it is to measure localization and separation cues and hence it is more difficult to recognize the acoustic sources of interest. The objective of this research is to investigate signal processing and machine learning techniques that can be used for noise-robust acoustic target recognition using miniature microphone arrays.

The first part of this research focuses on designing "smart" analog-to-digital conversion (ADC) algorithms that can enhance acoustic cues in sub-wavelength microphone arrays. Many source separation algorithms fail to deliver robust performance when applied to signals recorded using high-density sensor arrays where the distance between sensor elements is much less than the wavelength of the signals. This can be attributed to limited dynamic range (determined by analog-to-digital conversion) of the sensor which is insufficient to overcome the artifacts due to large cross-channel redundancy, non-homogeneous mixing and high-dimensionality of the signal space. We propose a novel framework that overcomes these limitations by integrating statistical learning directly with the signal measurement (analog-to-digital) process which enables high fidelity separation of linear instantaneous mixture. At the core of the proposed ADC approach is a min-max optimization of a regularized objective function that yields a sequence of quantized parameters which asymptotically

tracks the statistics of the input signal. Experiments with synthetic and real recordings demonstrate consistent performance improvements when the proposed approach is used as the analog-to-digital front-end to conventional source separation algorithms.

The second part of this research focuses on investigating a novel speech feature extraction algorithm that can recognize auditory targets (keywords and speakers) using noisy recordings. The features known as Sparse Auditory Reproducing Kernel (SPARK) coefficients are extracted under the hypothesis that the noise-robust information in speech signal is embedded in a subspace spanned by sparse, regularized, over-complete, non-linear, and phase-shifted gammatone basis functions. The feature extraction algorithm involves computing kernel functions between the speech data and pre-computed set of phased-shifted gammatone functions, followed by a simple pooling technique ("MAX" operation). In this work, we present experimental results for a hidden Markov model (HMM) based speech recognition system whose performance has been evaluated on a standard AURORA 2 dataset. The results demonstrate that the SPARK features deliver significant and consistent improvements in recognition accuracy over the standard ETSI STQ WI007 DSR benchmark features. We have also verified the noise-robustness of the SPARK features for the task of speaker verification. Experimental results based on the NIST SRE 2003 dataset show significant improvements when compared to a standard Mel-frequency cepstral coefficients (MFCCs) based benchmark.

I dedicate this dissertation to my parents, for their love and support.

## ACKNOWLEDGMENT

I would like to take this opportunity to acknowledge several people who have been particularly inspiring and helpful to my research work.

First I would like to thank my advisor, Dr. Shantanu Chakrabartty for the tremendous time, energy, and wisdom he invested in my Ph.D. education.

I would like to thank the other members of my Ph.D. thesis committee, Prof. Hayder Radha, Prof. Lalita Udpa, and Dr. Rong Jin, for their valuable feedback, suggestions, and time that helped me improve the quality of this work.

I would like to thank my past and current colleagues at AIM lab, who shared their friendship with me: Amit Gore, Yang Liu, Chenling Huang, and Ming Gu.

Finally, there is no way I would be where I am today without the immeasurable love, support, and encouragement of my parents Manouchehr Fazel and Behjat Kazemi and my lovely fiancee Soodabeh.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1

# Introduction

One of several emerging areas where micro/nano scale integration promises significant breakthroughs is in the field of acoustic sensing, separation, and recognition. For example, it is envisioned that next generation of intelligent hearing devices will integrate hundreds of micro/nanoscale microphones, separate speech from noise, track conversations in cluttered environments and thus provide significant improvements in speech intelligibility for individuals with hearing impairments. Sensing, separation and recognition of acoustic sources using micro/nano scale microphone arrays, however, poses significant challenges in the area of robust signal processing. The objective of this research is to develop theory and algorithm for robust acoustic recognition systems using miniature microphone arrays and to investigate using of these devices in real world applications.

## 1.1 Motivations and applications

The acoustic sensing and recognition has been widely used in different applications ranging from bioacoustics to military devices. In bioacoustics [1, 2, 3, 4], acoustic sensing and recognition systems have been used by ornithologists to study bird species interaction in their environment.

The acoustic based technology in bioacoustic is particularly important in places where the visibility is limited such as rain forest environment. In military applications [5, 6, 7, 8, 9, 10], acoustic sensing and detection system will detect an acoustic event, such as a sniper's weapon firing or a door slamming and then will use that information for further actions. These acoustic sensing/detection devices are usually mounted on a robotic vehicles which provides commanders with overall situational awareness. Target detection and tracking systems also partially benefit from the acoustic sensing technology [11, 12]. The acoustic sensing system has also been used in electronic textiles (e-textiles) with applications mostly in military equipments [13, 14]. Acoustic sensing and recognition systems have also been utilized in intelligent transportation technology for different purposes such as speed monitoring, traffic counting, and vehicle detection and classification [15, 16]. Railroad system has also been benefited in using acoustic devices for bearing health monitoring [17, 18].

Micro/nano-scale acoustic recognition system is one of emerging areas in miniaturization techniques where they have found different applications across disciplines. There are some applications in which their imposed limitations motivate the use of miniature acoustic recognition systems. Some of these applications like hearing aids require the source of interest to be far away from the recording device. In such cases, it is very beneficial to use array of microphones in order to use the spatial information for environmental noise compensation. However, acoustic sensing in miniature microphone arrays introduces a key challenge which is the "high fidelity" imaging of the acoustic events in their surrounding environment. Other challenge comes from the acoustic recognition system where the objective is to robustly recognize the acoustic events in real environment which has been attracting a bulk of research in signal processing society. This research addresses these challenges and is particularly interested in micro/nano scale acoustic source separation and recog-

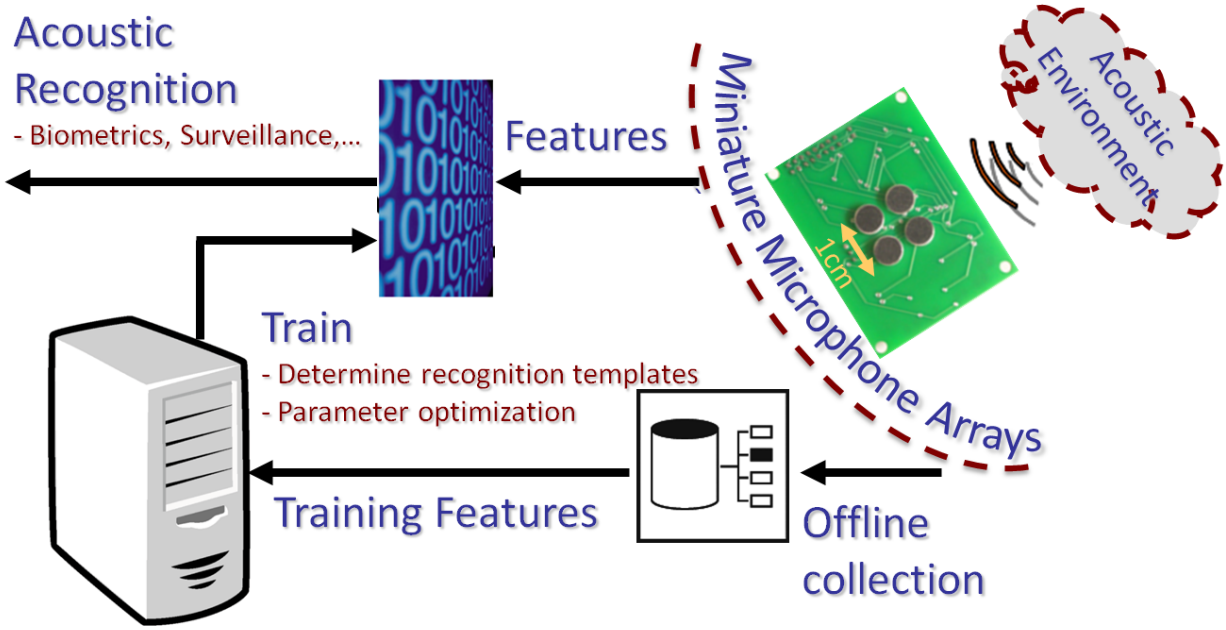



Figure 1.1: Motivation: Offline data collection and acoustic recognition (For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation)

nition systems where the multiple recording elements are places very close to each other in which the recording condition can be viewed as far-field.

Fig. 1.1 shows a big picture motivation behind this research. In such systems, an offline data collection provides the training data for the recognition system. This data is collected using the miniature microphone array and in an environment where it has as less mismatch as possible with the on-line recognition situation. This data then will be used to build an acoustic model. In the on-line acoustic recognition, first the acoustic event will be captured again using the miniature microphone array and then from the recorded data some appropriate feature will be extracted. Using the acoustic model generated offline, an acoustic target will be recognized from the extracted features. Usually the acoustic model in the offline training procedure, is built with the same features as the on-line recognition system. To motivate even more this research two specific applications of

miniature acoustic sensing and recognition is presented below.

**Biometric systems:** One of the technology were miniature microphone array promises breakthrough is in the area of speech based biometric systems. Speech based biometric system such as speaker verification/identification is a popular biometric identification technique used for authenticating and monitoring human subjects using their speech signal [19, 20]. The method is attractive because it does not require direct contact with the individual, thus avoiding the hurdle of "perceived invasiveness" inherent in most biometric systems like iris and finger print recognition systems. To date, most commercial implementations of speaker verification systems have been designed for enterprise applications that utilize large scale computing resources and infrastructure [21, 22, 24, 23]. However, with the proliferation of portable devices there has been an increasing demand for small scale speaker verification systems and therefore these systems demand recognition performance which is robust to variable background noise and to channel (microphone) conditions. Homeland security and surveillance applications might require identification/verification of target speakers in a given environment, in which case the speaker may not be in proximity with the microphone. The enrollment data corresponding to the target speaker could be limited and could have been acquired from unconventional sources (video/audio tapes, over the network or from archives). Unfortunately, even though most existing speech based recognition systems deliver acceptable accuracy under controlled/clean conditions, their performance degrades significantly when subjected to noise present in practical environments and especially for applications where the speaker is not in the proximity of the recording elements [25, 26, 27]. This shows that acquisition and recognition of the speech signal for small scale speech based biometric system needs high resolution signal and a robust recognition system where this research addresses these problems by employing the techniques for both super-resolution recording and robust recognition.

**Hearing aids:** Typically hearing aid users which were only $\%20$ of all the hearing impaired in 2004 [28], have difficulty when listening to a speaker in the noisy environment. This difficulty comes from the fact that the conventional hearing aids amplify all receiving sounds without discriminating between the speaker of the interest and background noises. The ideal hearing aid is the one with the functionality of binaural hearing system which provides the signals and allows the patient to listen to one speaker. The ability to record the high quality speech signal using microphone arrays and the lack of success on single microphone hearing aids motivate microphone array hearing aids [29, 30, 31]. Using the microphone arrays for hearing aids provides the special information of the receiving signals which can be used to focus on a specific speaker in the surrounding environment. The bulk of research concerning the speech array processing has been done [34]. Although the basic principle of speech array processing can be applied for hearing aid applications, several other problems exist which needs to be answered. One of them is the cosmetic consideration which limits the design of hearing aids to use the miniaturize/compact microphone arrays. Using miniature/compact microphone arrays makes the spatial aperture much smaller relative to the wavelength of the speech signal. Other problem is the noisy environment which exists for the operation of the hearing aid. Although this work does not directly address the specific hearing aids application needs, but the techniques suggested here can be used in the next generation of "smart" hearing aids.

System miniaturization is not only limited to the acoustic sensing applications but it has found applications in other areas as well. For example in the rapidly emerging field of brain machine interface, it is very common to record from thousands of neurons using micro-electrode sensor array [79, 80, 81]. The recording signals are then processed to extract useful information for controlling the movement of a prosthetic device. Another example comes from the area of wireless

communication where a 16-element transmitter array is used with a sub-wavelength inter-element spacing between sensors [35].

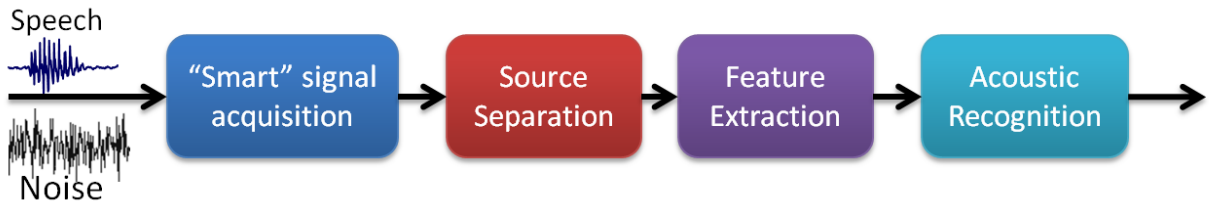



Figure 1.2: Acoustic recognition system composed of four main sub-systems

## 1.2 Miniature acoustic recognition system

In a micro/nano scale setting, this research proposes an acoustic recognition system composed of four main components: (a) “smart” signal acquisition, (b) source separation, (c) feature extraction, and (d) acoustic recognition where Fig. 1.2 shows a block diagram of the proposed system.

**The “smart” signal acquisition** unit is used to record the acoustic signals with high fidelity by using as much dynamic range as possible in analog-to-digital conversion module. This unit uses a miniature microphone array in order to be able to record signal of interest with high fidelity. Note that in many of applications where a miniature acoustic recognition system can be used, the signal of interest is far away from the recording elements, hence the need for a microphone array. The proposed signal acquisition device performs the spatial sensing of signal along with analong to digital process where we have formulated the analog-to-digital modulation within the framework of statistical learning such that the algorithm retrives the spatial manifold which contains the information for decorrelating the signal wavefront. In the current research, a min-max optimization approach is used to model the signal de-correlation and analog-to-digital conversion with a single cost function. In order to optimize the cost function, a stochastic gradient descent

and ascent algorithms are employed. The stochastic gradient descent is used to minimize the cost function with respect to the internal state of the system in which it yeilds the analog-to-digital conversion module. The stochastic gradient ascent maximizes the cost function with respect to the signal transformation that minimizes the input correlation hence decorrelating the input signals. The architecture of the "smart" signal acquisition is shown in Fig. 1.3 where the input is a time varying analog signal $\mathbf{x}$. This system consists of a matrix-vectormultiplier which transforms the input signal $\mathbf{x}$ into $\mathbf{Bx}$ where $\mathbf{B}$ denotes a linear transformation matrix. This transformed signal is then processed by an array of analog-to-digital converter to produce a binary data stream $\mathbf{d}$ along with the spatial information $\mathbf{\Lambda}$ and $\mathbf{B}$. An adaptation unit uses the binary output $\mathbf{d}$ to update the $\mathbf{B}$ and $\mathbf{\Lambda}$.

The source separation system is used in order to separate the source of interest from the rest of acoustic events and to provide a high quality speech signal for the recognition system. The assumption here is that the input speech signals are statistically independent from each othre, therefore conventional independent component analysis (ICA) algorithms were applied to separate the sources. After separating the source of interest from mixutre of different signals, the next step is to extract the robust features for the recognition.

**The feature extraction** unit provides speech features that are robust to the environmental noises. A hierarchical model is used to extract the robust auditory features from the speech signal. This model is based on the recent finding in auditory neuroscience indicating that there is a hierarchical processing in the human auditory cortex where the received signal is first broken down into basic features and later they are integrated into more complex stimuli. Inspiring from biological data, the proposed hierarchical model consists of two layers of processing as shown in Fig. 1.4. In the first level of this computational model, the similarity of sensory auditory world is measured

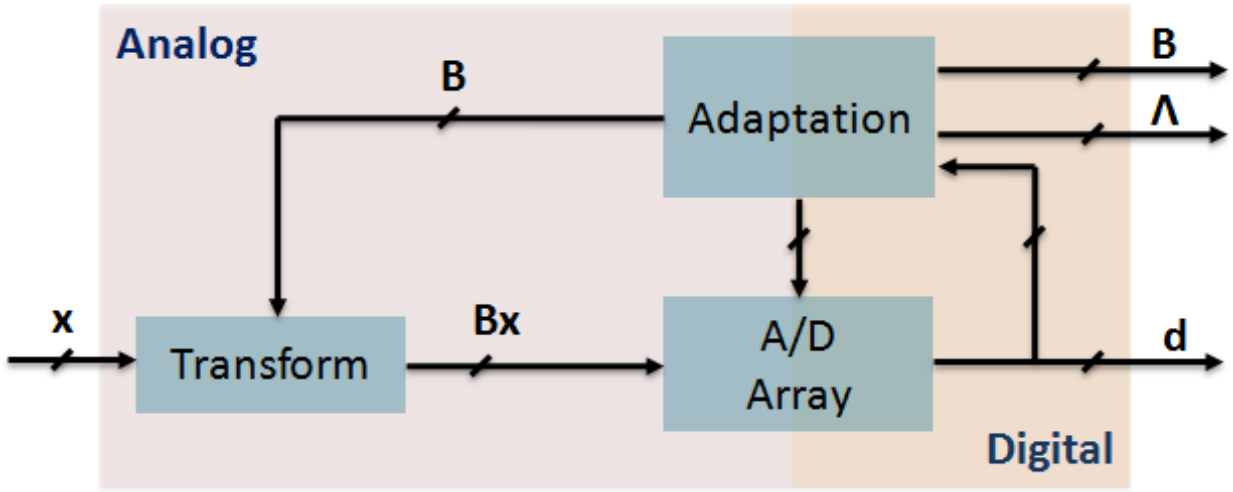



Figure 1.3: Architecture of the "smart" signal acquisition device

through a kernel based approach with a set of gammatone basis functions. These simple basis functions represent the so-called spectro-temporal receptive fields (STRFs) in the auditory cortex. In order to implement this, we apply the kernel based approach to a reproducing kernel Hilbert space (RKHS) spanned by gammatone basis functions. The result of incorporating this a-priori information is that these signitures can be extracted in real-time using pre-computed projection matrices. Then all the outputs are sent to the higher level where they will be integrated in order to generate the more complex outputs. In current research, we explored two different idealized pooling mechanisms of summation ("SUM") and maximization ("MAX") operation, both with nonlinear weights to integrate the outputs of previous level. This weighting function also emulates the psychoacoustical nonlinear relation between the intensity of sound and its perceived loudness. The proposed computational model is very close to the HMAX approach introdued in [63] where it models the visual cortex in a hierarchical fashion for objective recognition task. In order to feed these features into a back-end acoustic recognition system, discrete cosine transform (DCT) is used to decorrelate the features.

**The acoustic recognition** unit is used to recognize the acoustic events. Once the feature vectors

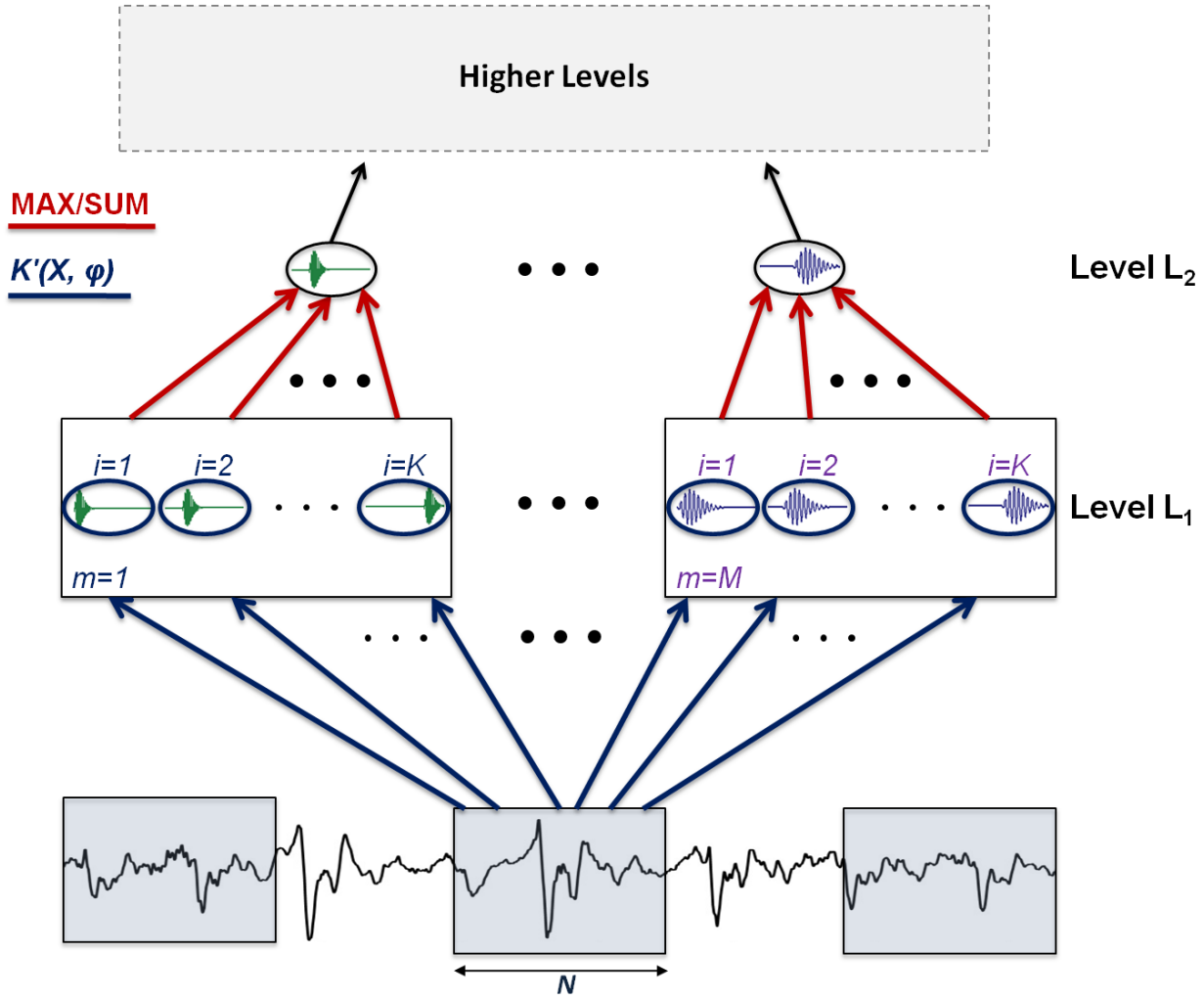



Figure 1.4: Hierarchical model of auditory feature extraction.

corresponding to the acoustic events have been extracted the associated data also known as training data is used to build models for the recognition systems in an offline process. During the test phase, the trained models are used to recognize a sequence of feature vectors extracted from unknown acoustic events.

Even the proposed miniature acoustic recogniton system is general and can be used for recognition of any type of acoustic events, but in this research we used it for two tasks of speech recognition and speaker verification.

## 1.3 Scientific contributions

The conducted research has the following two main scientific contributions:

**"Super-resolution" high-density acoustic signal acquisition:** In this research a far-field recording condition on miniature microphone array has been investigated, a research area that have not been received much attention. First a mathematical model has been developed for the miniature microphone array. This model shows that speech signals received at the miniature microphone array can be considered to be in the far-field condition with the instantaneous mixing. Then a "smart" signal acquisition system is introduced in order to remove the correlation of the received signal at the analog-to-digital conversion level and increasing the dynamic range of the acquisition system. This method is based on a min-max optimization of a regularized objective function which integrates the analog-to-digital conversion with the statistical machine learning.

**Hierarchical auditory features:** This research also proposes a novel speech feature extraction method based on a hierarchical fashion to improve the robustness of the acoustic recognition system by exploiting properties of a functional regression procedure in a reproducing kernel Hilbert space (RKHS) [64, 67, 65, 66]. This method is based on the hypothesis that robustness in speech signal is encoded in temporal and spectral manifolds (represented here by kernels) which remain intact even in presence of ambient noise. However, under clean recording conditions (laboratory setting), most learning algorithms like hidden Markov models (HMMs) [104] and support vector machines (SVMs) [68] exploit only linear dominant features which unfortunately can easily be corrupted by ambient noise. RKHS regression endows the proposed innovative features with the following robustness properties:

- The algorithm doesn't make any prior assumption on noise statistics.
- The algorithm uses kernel methods to extract features that are nonlinear and robust to cor-

ruption by noise.

- Robust parameter estimation is ensured by imposing smoothness constraints based on regularization principles.

- The proposed signitures can be extracted in real-time using pre-computed projection matrices.

## 1.4 Dissertation organization

The dissertation is organized as follows: Chapter 2 motivates the "smart" audio signal acquisition systems as substitutions for conventional Nyquist ADCs for miniature microphone arrays. Then a mathematical model for signal acquisition in miniature/compact microphone array is presented. The model shows that the signal recorded from miniature array is near singular and conventional ways of signal acquisition fail to deliver a robust performance due to limited dynamic range of microphone which is determined by analog-to-digital conversion. This limitation is coming from the fact that a large cross-channel redundancy and non-homogeneous mixing is presented in recorded signal space on miniature microphone array. The proposed "smart" signal acquisition device is then presented in chapter 3. The core of this technique is based on a min-max optimization framework that can efficiently and adaptively quantize non-redundant analog signal sub-spaces which leads to significant performance improvement for any DSP based source separation algorithm. The performance of the proposed signal acquisition device is evaluated using synthetic and real recordings. Experiments have been shown to demonstrate significant and consistent performance improvements when the proposed approach is used as the analog-to-digital front-end to conventional source separation algorithms. A detail overview of the acoustic signal recognition system

is presented in chapter 4. In this chapter a brief introduction to statistical pattern recognition techniques that are commonly used in acoustic recognition will be provided. This includes an overview of some of the basic functional units such as speech feature extraction, acoustic modeling, and classification. Then this chapter discusses some of the commonly used techniques which make real-world acoustic recognition systems more robust in noisy environment. Chapter 5 introduces a novel hierarchical model to extract the auditory speech features. This model uses a regression technique in a reproducing kernel Hilbert space (RKHS) in order to measure the similarity of sensory auditory world. In this chapter, the theory behind these features that are known as Sparse Auditory Reproducing Kernel (SPARK) is first described. They are extracted under the hypothesis that the noise-robust information in speech signal is embedded in a subspace spanned by overcomplete, regularized and normalized gammatone basis functions. In the last part of this chapter two benchmarks is presented for acoustic recognition systems: the first one is a HMM based speech recognition system and the second one is an SVM-based speaker verification system. Using these benchmarks, the performance of the proposed system is evaluated and compared to the conventional acoustic recognition systems. Concluding remarks and the future directions for the presented work are discussed in chapter 6.

# Chapter 2

# Smart Audio Signal Acquisition Devices

## 2.1 Motivation for smart audio signal acquisition devices

Miniature microphone arrays for sensing the acoustic events are becoming more common for different applications. One of such applications is an acoustic recognition system where the objective is to identify a person as in a speaker recognition system or convert the speech into text as in a speech recognition system. However using micro/nano-scale microphone arrays in an acoustic recognition system poses a key challenge to image acoustic events occurring in its environment with high fidelity (spatial and temporal) where due to the dispersive nature of the surrounding media, each element of the sensor array records a mixture of signals generated by the source of interest and other events (noises) in its environment. In order to recognize the source of interest, the proposed recognition system takes advantage of a source separator. However several factors limit the performance of traditional source separation techniques and hence the performance of the acoustic recognition system when acoustic signal acquired from miniature/compact microphone arrays:

- **Far-field effects:** For miniature arrays, sources are usually located at distances much larger than the distance between recording elements. As a result, the mixing of signals at the recording elements is near singular. Recognition of the acoustic source of interest needs separation of that source from near ill-conditioned mixtures which would require super-resolution signal processing to reliably identify the parameters of the separation manifold.

- **Near-far effects:** For miniature sensor arrays, a stronger source that is closer to the array can mask the signal produced by background sources. Separating the background sources in presence of the strong masker would again require super-resolution processing of the input signals.

DSP based source separation algorithms are typically implemented subsequent to a quantization operation (analog-to-digital conversion) and hence do not consider the detrimental effects of finite resolution due to the quantizer. In particular, for a high-density sensor array, a naive implementation of a quantizer that uniformly partitions each dimension (pulse coded modulation) of the input signal space could lead to a significant loss of information. To understand the effect of this degradation consider the framework of a conventional source separation algorithm as shown in Fig. 2.1(a). The "analog" signal $\mathbf{x} \in \mathcal{R}^M$ recorded at each of the sensor array is given by $\mathbf{x} = \mathbf{A}\mathbf{s}$, $\mathbf{A} \in \mathcal{R}^M \times \mathcal{R}^M$ being the mixing matrix and $\mathbf{s} \in \mathcal{R}^M$ being the independent sources of interest. This simplified linear model is applicable to both instantaneous mixing as well as to convolutive mixing formulation [37]. The recorded signals are first digitized and then processed by a digital signal processor (DSP) which implements the source separation algorithm. For the sake of simplicity, assume that the algorithm is able to identify the correct unmixing matrix given by $\mathbf{W} = \mathbf{A}^{-1}$ which is then used to recover the source signals $\tilde{\mathbf{s}} \in \mathcal{R}^M$. The effect of quantization

in this approach can be captured using a simple additive model as shown in Fig. 2.1(a)

$$\tilde{\mathbf{s}}_d = \mathbf{W}(\mathbf{x} + \mathbf{q}) = \mathbf{s} + \mathbf{A}^{-1}\mathbf{q} \tag{2.1}$$

where $\mathbf{q}$ denotes the additive quantization error introduced during the digitization process. The reconstruction error between the recovered signal $\tilde{\mathbf{s}}_d$ and the source signal $\mathbf{s}$ can then be expressed as

$$||\tilde{\mathbf{s}}_d - \mathbf{s}|| = ||\mathbf{A}^{-1}\mathbf{q}|| \leq ||\mathbf{A}^{-1}||.||\mathbf{q}|| \tag{2.2}$$

where $||.||$ denotes a matrix and vector norm [38]. Equation (2.2) indicates that under ideal reconstruction conditions, the performance of conventional source separation algorithm is limited by: (a) quantization error (accuracy of analog-to-digital conversion) and (b) the nature of the mixing operation determined by $\mathbf{A}$. For high-density sensors, the mixing typically tends to be ill-conditioned ($||\mathbf{A}^{-1}|| \gg 1$), as a result the reconstruction error due to equation (2.2) could be large.

Now consider the framework shown in Fig.2.1(b) which is at the core of the proposed resolution enhancement approach. The signals recorded by the sensor array is first transformed by $\mathbf{P}$ (in the analog domain) before being quantized. In this case, the reconstructed signal $\tilde{\mathbf{s}}_m \in \mathcal{R}^M$ can be expressed as

$$\tilde{\mathbf{s}}_m = \mathbf{D}(\mathbf{P}\mathbf{x} + \mathbf{q}). \tag{2.3}$$

For source separation $\mathbf{DP} = \mathbf{A}^{-1}$, for which the reconstructed signal now can be expressed as

$$\tilde{\mathbf{s}}_m = \mathbf{s} + \mathbf{A}^{-1}\mathbf{P}^{-1}\mathbf{q} \tag{2.4}$$

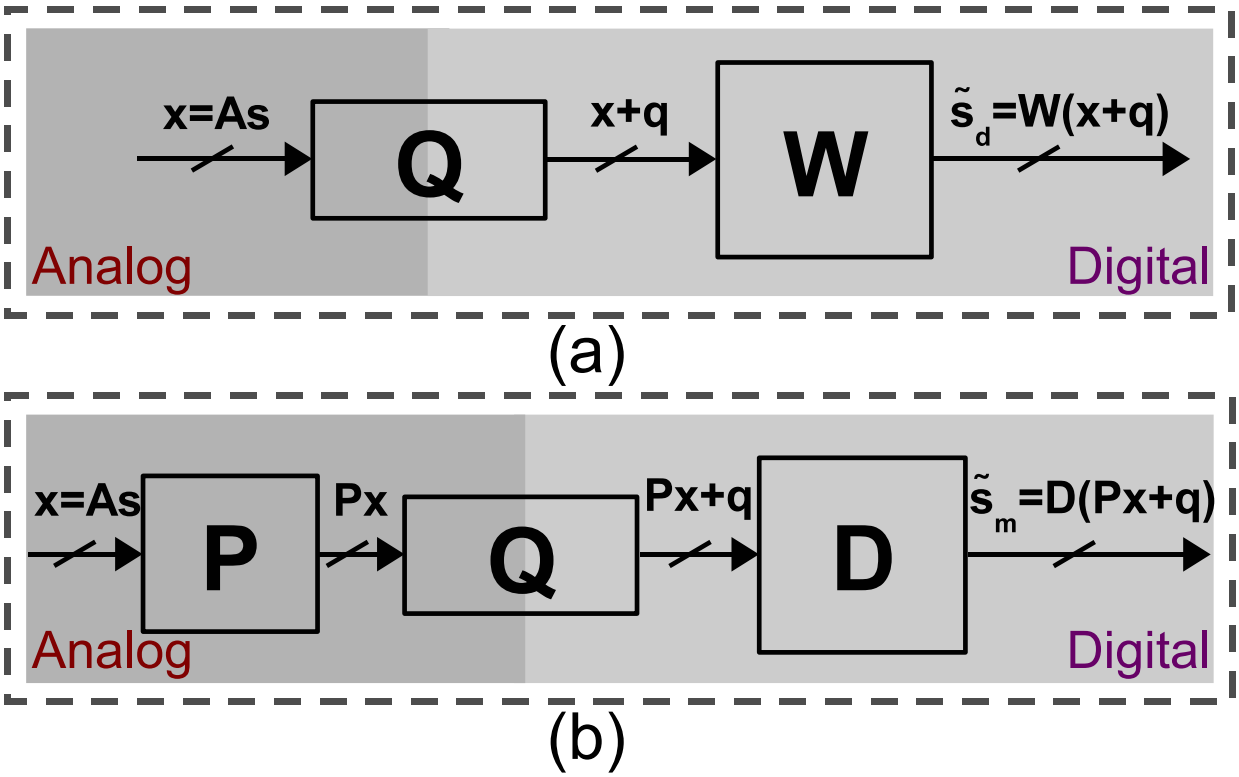



Figure 2.1: System architecture where the source separation algorithm is applied (a) after quantization (b) after analog projection and quantization

which leads to the reconstruction error

$$||\tilde{\mathbf{s}}_m - \mathbf{s}|| = ||\mathbf{A}^{-1}\mathbf{P}^{-1}\mathbf{q}|| \leq ||\mathbf{A}^{-1}\mathbf{P}^{-1}||.||\mathbf{q}||. \tag{2.5}$$

Thus, the reconstruction error can now be controlled by the choice of the transform $\mathbf{P}$ and is not completely determined by the mixing transform $\mathbf{A}$. An interesting choice of the matrix $\mathbf{P}$ is the one that satisfies

$$||\mathbf{A}^{-1}\mathbf{P}^{-1}|| = 1 \tag{2.6}$$

which ensures that the input signals are normalized before processed by the DSP based source separation algorithm. Equation 2.5 then reduces to

$$||\tilde{\mathbf{s}}_m - \mathbf{s}|| \leq ||\mathbf{q}|| \tag{2.7}$$

and the expected performance improvement when employing the framework in Fig. 2.1(b) over the framework in Fig. 2.1(a) is given by

$$PI = -20 \log ||\mathbf{A}^{-1}||. \tag{2.8}$$

Equation (2.8), thus, shows that the for near-singular mixing, $||\mathbf{A}^{-1}|| \gg 1$, the performance improvement based on the resolution enhancement technique shown in Fig. 2.1(b) could be significant. However, the performance improvement is valid only if the analog projection $\mathbf{P}$ can be precisely and adaptively determined during the process of quantization ("analog-to-digital" conversion). This procedure is unlike traditional multi-channel "analog-to-digital" conversion where each signal channel is uniformly quantized the input signal without taking into consideration the spatial statistics of the input signal. Since the projection $P$ is also quantized, the precision to which the condition (2.6) is satisfied is also important. In this regard, oversampling "analog-to-digital" converters like $\Sigma\Delta$ modulators are attractive since the topology is robust to analog imperfection and can easily achieve dynamic ranges greater than 120 dB (more than 16 bits or accuracy) [39]. In this research we show that the learning algorithm can efficiently and adaptively quantize non-redundant analog signal sub-spaces which leads to significant performance improvement for any DSP based source separation algorithm and hence the proposed acoustic recognition system. This innovative approach which is called "super-resolution Sigma-delta" will be discussed in chapter 3.

## 2.2 Signal acquisition in miniature microphone array

In this section a mathematical model for a miniature microphone array is presented. The model shows that the recorded signals from the array can be near singular. Then this model will be used later to show the superior performance of proposed smart signal acquisition process over the standard ADC. This modeling resort to the far-field wave propagation models that have been extensively studied within the context of array processing and *plenacoustic* models [82, 83, 84]. Knowing the *plenacoustic function*, the actual sound at a desired position in a sound field especially in a room can be modeled via the convolution of this function with the source signal. For the modeling purpose, consider a microphone array shown in Fig. 2.2 that consists of two recording elements. If the inter-element distance is much less than the wavelength of the microphone signal of interest, the signals recorded at each of the sensor elements can be approximated using far-field models. For example, for audio signals (100-20,000 Hz), a far-field model can be assumed for microphone arrays with inter-element distances less than $3.4cm$ (coherence length). Also, for miniature microphone array the distance to the sources from the center of the array can be assumed to be larger than the inter-element distance. We express the signal $x_j(\mathbf{p}_j, t)$ recorded at $j^{th}$ microphone as a superposition of $i$ independent sources $s_i(t)$ ($i \in 1, .., D$), each of which are referenced with respect to the center of the array [83]. This can be written as

$$x_j(\mathbf{p}_j, t) = \sum_i c_i(\mathbf{p}_j) s_i(t - \tau_i(\mathbf{p}_j)) \tag{2.9}$$

where $c_i(\mathbf{p}_j)$ and $\tau_i(\mathbf{p_j})$ denotes the attenuation and delay, for the source $s_i(t)$ at the position $\mathbf{p}_j$, measured relative to the center of the sensor array. $\mathbf{p}_j$ in equation (2.9) denotes the position vector of the $j^{th}$ microphone. Equation (2.9) can be approximated using Taylor's series expansion as

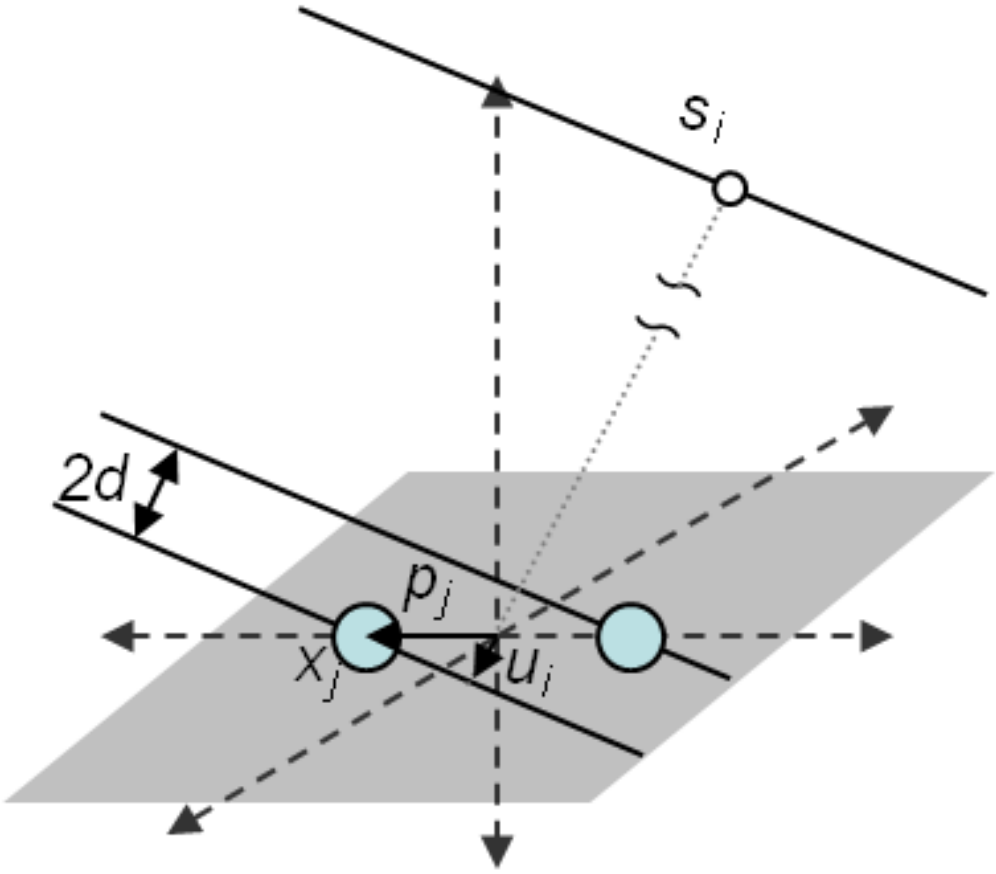



Figure 2.2: Far-field recording on a miniature microphone arrays.

$$x_j(\mathbf{p}_j, t) = \sum_i c_i(\mathbf{p}_j) \sum_{k=0}^{\infty} \frac{(-\tau_i(\mathbf{p}_j))^k}{k!} s_i^{(k)}(t) \tag{2.10}$$

Under far-field conditions it can be assumed that $c_i(\mathbf{p}_j) \approx c_i$ is constant across all the sensor elements. Also, the higher-order terms in the series expansion (2.10) can be ignored and can be expressed as

$$x_j(\mathbf{p}_j, t) \approx \sum_i c_i s_i(t) - \sum_i c_i \tau_i(\mathbf{p}_j) \dot{s}_i(t). \tag{2.11}$$

The component $x_C(t) = \sum_i c_i s_i(t)$ signifies the common-mode signal common to all the recording elements and the second part of the RHS signifies an instantaneous mixture of the derivative of the source signals. The common-mode component can be canceled using a differential measure-

ment [85] under which equation (2.11) becomes

$$\Delta x_j(\mathbf{p}_j, t) = x_j(\mathbf{p}_j, t) - x_C(t) \approx -\sum_i c_i \tau_i(\mathbf{p}_j) \dot{s}_i(t) \tag{2.12}$$

and can be expressed in a matrix-vector form as

$$\Delta \mathbf{x}(\mathbf{t}) \approx -\mathbf{A}\dot{\mathbf{s}}(\mathbf{t}) \tag{2.13}$$

where $\mathbf{A} = \{c_i \tau_i(\mathbf{p}_j)\}$ denotes the instantaneous mixing matrix. Under far-field approximation, the time delays can be expressed as

$$\tau_i(\mathbf{p}_j) = \mathbf{u}_i^T \mathbf{p}_j / v \tag{2.14}$$

where $\mathbf{u}_i$ is the unit normal vector of the wavefront of source $i$ and $v$ is the speed of sound ($v = 340m/s$ in air). Thus equations (2.13) and (2.14) show that for miniature recording array, recovery of the desired sources $\mathbf{s}$ or $\dot{\mathbf{s}}$ entails solving a linear source separation problem [86]. However, equation (2.14) reveals that sources that are located closer to the sensor array can completely mask the sources located away from the sensor array, resulting in a near-singular mixing $\mathbf{A}$. As shown in the following description that under near-singular mixing conditions, conventional methods of signal acquisition and source separation fail to deliver robust performance.

Many source separation and speech feature extraction algorithms fail to deliver robust performance when applied to signals recorded using miniature microphone arrays. This can be a result of limited dynamic range (determined by analog-to-digital conversion) of the microphone which is insufficient to overcome the artifacts due to large cross-channel correlation, non-homogeneous

mixing and high-dimensionality of the signal space. In the next chapter a novel framework will be proposed that overcomes these limitations by integrating statistical learning directly with the signal measurement (analog-to-digital) process which enables high fidelity separation of linear instantaneous mixtures which we saw it in this section for miniature microphone array.

# Chapter 3

# Sigma-Delta Learning

The underlying principle behind the proposed technique is illustrated using Fig. 3.1 which shows a two dimensional signal distribution along with the respective signal quantization levels (depicted using rectangular tick marks). In this example, the signal distribution has been chosen to cover only a small region of the quantization space which would be the case for near-singular mixing. Thus, in a traditional implementation where each dimension is quantized independently of the other there would be a significant information loss due to quantization. This approach towards estimating $\mathbf{P}$ (which was previously introduced in chapter 1) while performing signal quantization will be to decompose $\mathbf{P} \in \mathcal{R}^M \times \mathcal{R}^M$ as a product two simple matrices $\Lambda \in \mathcal{R}^M \times \mathcal{R}^M$ and $\mathbf{B} \in \mathcal{R}^M \times \mathcal{R}^M$ such that $\mathbf{P} = \Lambda\mathbf{B}$. The transformation matrix $\mathbf{B}$ will first "approximately" align the data distribution along the orthogonal axes, each axis representing an independent (orthogonal) component (shown in Fig. 3.1(b)). Based on this alignment, the signal distribution will be scaled according to a diagonal matrix $\Lambda$ such that the quantization levels now span a significant region of the signal space (Fig. 3.1(c)). Our objective will be to compute these transforms $\mathbf{B}$ and $\Lambda$ recursively while performing signal quantization. Even though the proposed procedure bears similarity

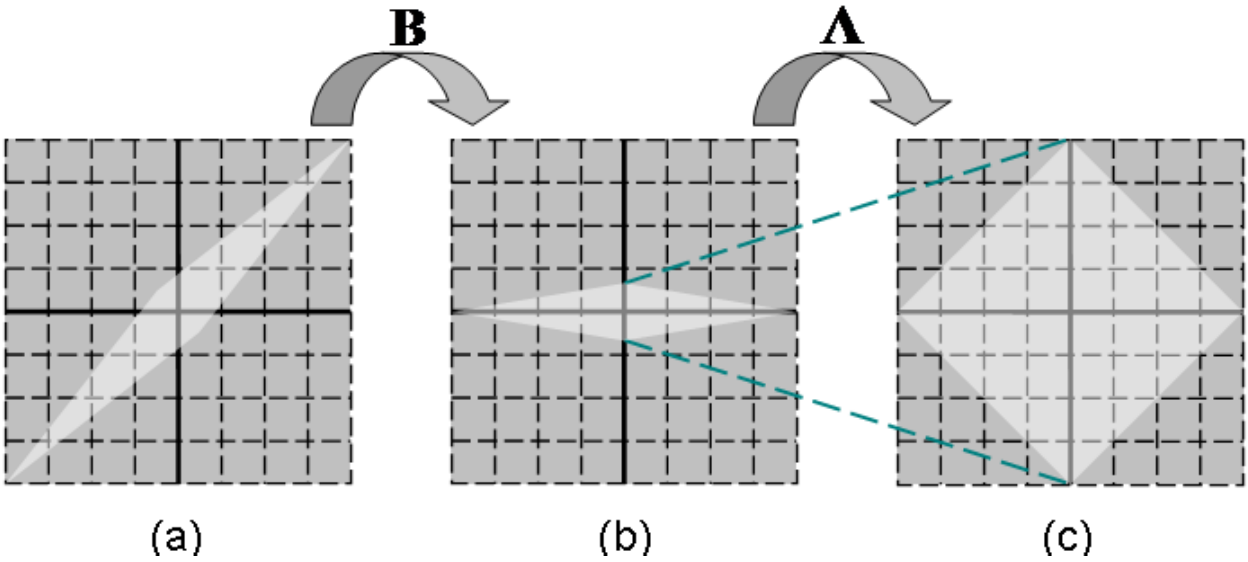



Figure 3.1: Illustration of the two-dimensional signal distribution for: (a) the input signals ; (b) signals obtained after transformation $\mathbf{B}$ and (c) signals obtained after resolution enhancement

to recursive techniques reported in many online "whitening" algorithms [87, 88, 89], the key difference for the proposed approach is that the adaptation and estimation of the projection matrix $P$ is coupled with the quantization process. Thus, unlike traditional online "whitening" techniques, in the proposed approach any imperfections or errors in the quantization process can be corrected through the adaptation of $P$. This approach can therefore be visualized as a "smart" analog-to-digital converter as shown in Fig. 3.2 which not only produces quantized (digitized) representation of the input signal $\mathbf{d}$ but also quantized (digitized) representations of the transforms $\mathbf{B}$ and $\Lambda$. In our formulation, the estimation algorithm for $P$ ($\mathbf{B}$ and $\Lambda$) has been integrated within a $\Sigma\Delta$ modulation algorithm and hence the name "$\Sigma\Delta$ learning. The choice of $\Sigma\Delta$ modulation is due to its robustness to hardware level artifacts (mismatch and non-linearity) which makes the modulation amenable for implementing high-resolution analog-to-digital converters [94]. Before presenting a generalized formulation for $\Sigma\Delta$ learning, first an optimization framework will be presented that can model the dynamics of first-order $\Sigma\Delta$ modulation.

## 3.1 Stochastic gradient decent and $\Sigma\Delta$ modulators

First, a one dimensional example of $\Sigma\Delta$ will be presented to illustrate how a $\Sigma\Delta$ modulator can be modeled as an equivalent stochastic gradient descent based optimization problem. Consider an architecture of a well known first-order $\Sigma\Delta$ modulator [94] as shown in Fig. 3.3(a) which consists of a single discrete-time integrator in a feedback loop. The loop also consists of a quantizer $Q$ which produces a sequence of digitized representation $d[n]$, where $n = 1, 2, ..$ denotes a discrete time-index. Let $x[n] \in \mathcal{R}$ be the sampled analog input to the modulator and without any loss of generality let $d[n] \in \{+1, -1\}$ be the output of a single-bit quantizer given by $d[n] = \text{sgn}(v[n-1])$ (Fig. 3.3(b)) where $v_n \in \mathcal{R}$ is the internal state variable or the output of the integrator as shown in Fig. 3.3(a). Then, the $\Sigma\Delta$ modulator in Fig. 3.3(a) implements the following recursion:

$$v[n] = v[n-1] + x[n] - d[n] \tag{3.1}$$

It can be seen from equation (3.1) that if $v[n]$ is bounded for all $n$, then

$$\frac{1}{N}\sum_{n=1}^{N} d[n] \overset{N\to\infty}{\longrightarrow} \frac{1}{N}\sum_{n=1}^{N} x[n]. \tag{3.2}$$

This implies that $\Sigma\Delta$ algorithm given by equation (3.1) produces a binary sequence $d[n]$ whose temporal average asymptotically converges to the temporal average of the input analog signal. This statistical dynamics is at the core of most $\Sigma\Delta$ modulators. However, from the perspective of statistical learning, the $\Sigma\Delta$ recursion in equation (3.1) can be viewed as a stochastic gradient step of the following optimization problem:

$$\min_{v} C(v) = \min_{v}[|v| - v\mathcal{E}_x(x)] \tag{3.3}$$

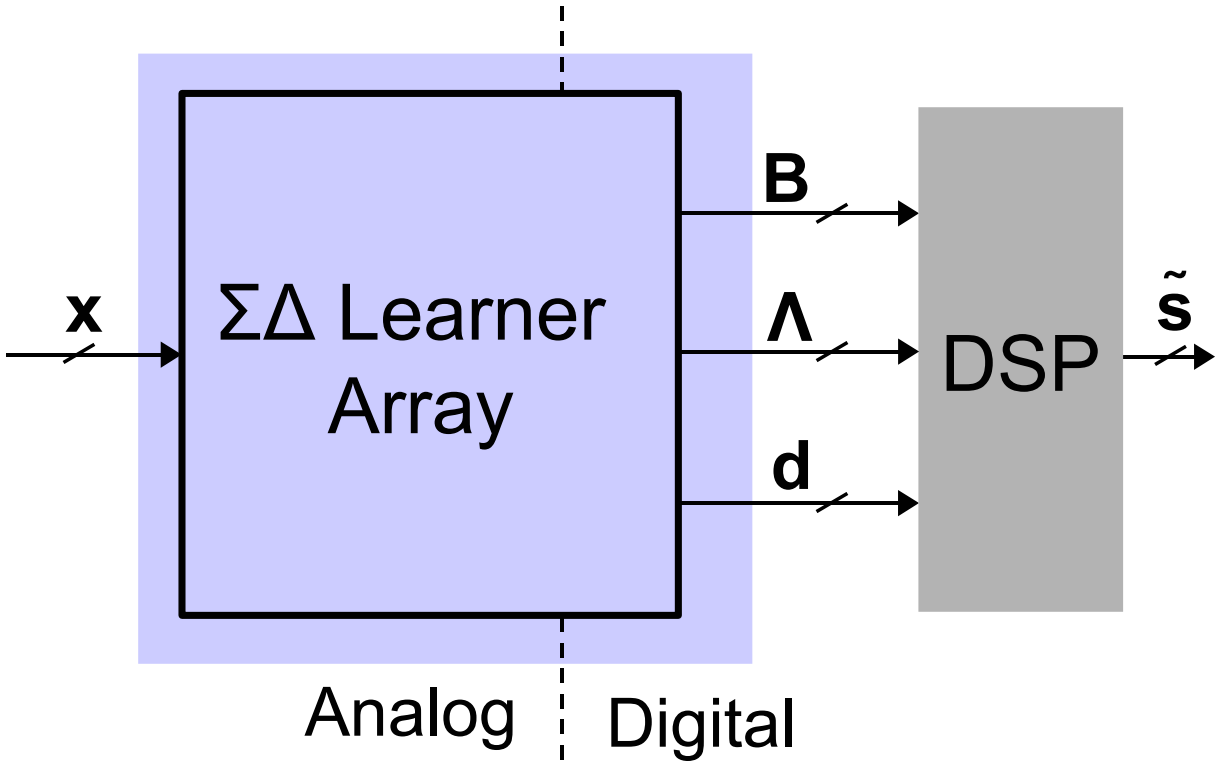



Figure 3.2: Architecture of the proposed sigma-delta learning applied to a source separation problem

where $\mathcal{E}_x(.)$ is the ensemble expectation of the random variable $x$. The optimization function $C(v)$ is shown in Fig. 3.3(c) for the case $|\mathcal{E}_x(x)| < 1$. The minima under this condition is $\min_v C(v) = 0$ which is achieved for $v = 0$ and thus does not contain any information about the statistical property of $x$. The recursion (3.1) ensures that $v[n]$ approaches the minima and then exhibits oscillations about the minima (shown in Fig. 3.3(c)). Note that unlike conventional stochastic gradient based optimization techniques [95], recursion (3.1) does not require any learning rate parameters. This is unique to the proposed optimization framework where the stochastic gradient descent is used to generate limit-cycles (oscillations) about the minima of the cost function $C(v)$. The only requirement in such a framework is that the assumption that the input signal is bounded which ensures that the limit-cycles are bounded. Under this condition, the statistics of the limit-cycles can asymptotically encode the statistics of the input signal with infinite precision, as shown by equation (3.2).

In the later sections, we will exploit this asymptotic property to precisely estimate the transform $\mathbf{P}$ which can then be used for resolving the acute spatial cues in miniature microphone arrays.

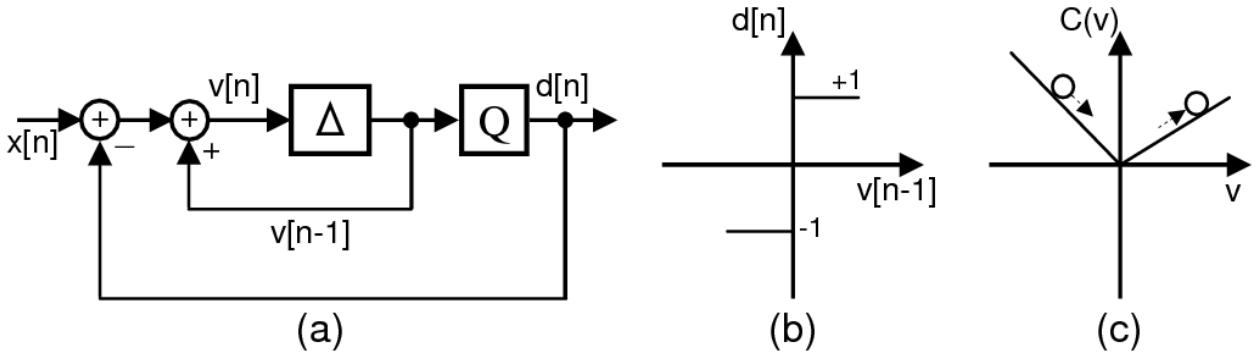



Figure 3.3: (a) System architecture of a first order $\Sigma\Delta$ modulator, (b) input-output response of single bit quantizer, and (c) illustration of "limit-cycle" oscillations about the minima of the cost function $C(v)$

Another unique aspect of the proposed optimization framework for modeling $\Sigma\Delta$ modulators is that the cost function $C(v)$ links "analog-to-digital" conversion through the regularizer $|v|$ whose derivative leads to a single-bit quantizer ($\mathrm{sgn}$ function). The second term in $C(v)$ ensures that the statistics of the quantized stream $d[n]$ matches the statistics of the input analog signal $x[n]$.

We now extend this optimization framework to a multi-dimensional $\Sigma\Delta$ modulator which uses a multi-bit quantizer and incorporates transformations $\mathbf{B}$ and $\Lambda$. Consider the following minimization problem

$$\min_{\mathbf{v}} \mathcal{C}(\mathbf{v}) \tag{3.4}$$

where the cost function $\mathcal{C}(\mathbf{v})$ is given by

$$\mathcal{C}(\mathbf{v}) = \Omega(\lambda^{-1}\mathbf{v}) - \mathbf{v}^T \mathcal{E}_x\{\mathbf{Bx}\}. \tag{3.5}$$

$\mathbf{x} \in \mathcal{R}^M$ is now an $M$ dimensional analog input vector and $\mathbf{v} \in \mathcal{R}^M$ is an internal state vector. For the first part of this formulation, $\lambda$ will be assumed to be a constant scalar and the transform $B$

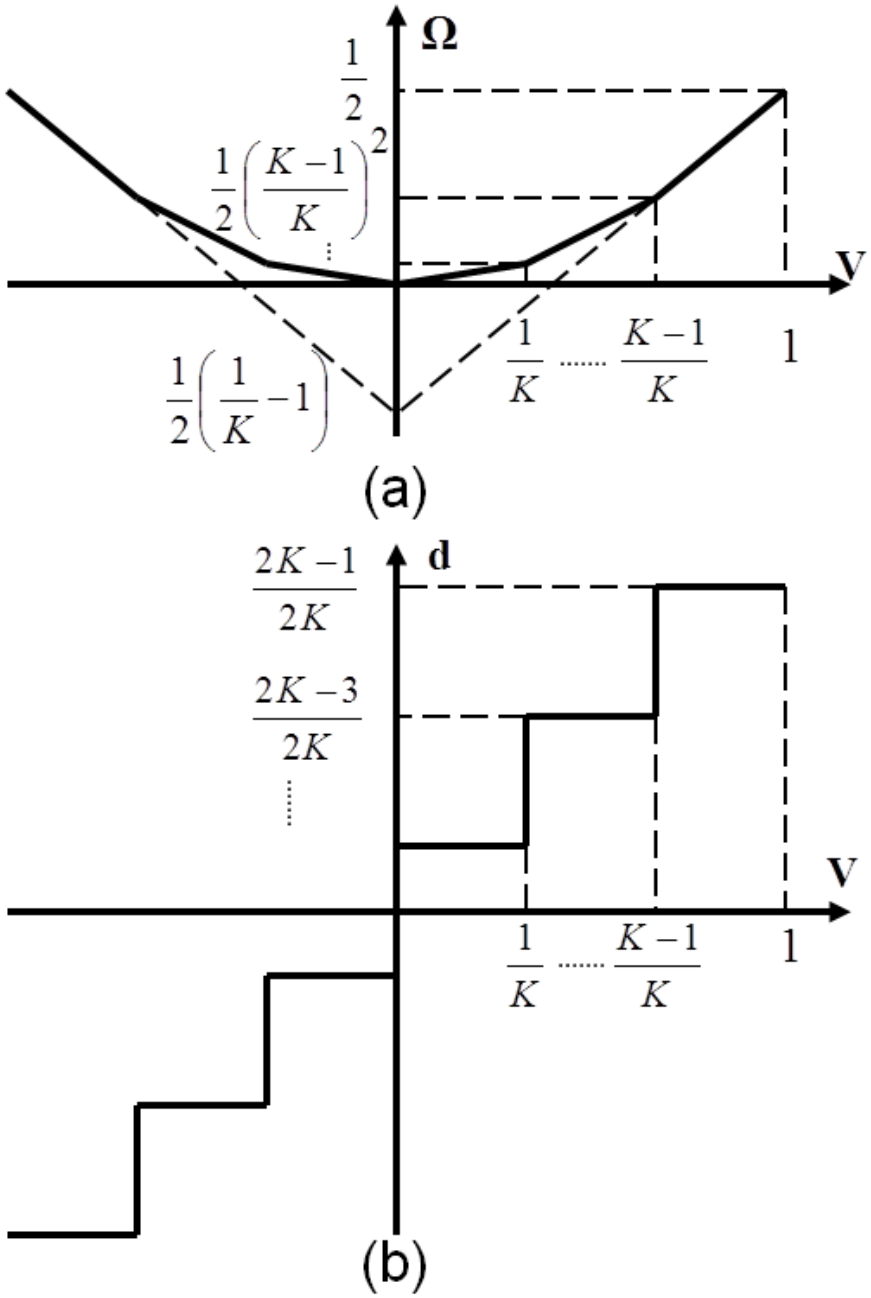



Figure 3.4: One dimensional piece-wise linear regularization functions and the multi-bit quantization function as its gradient

will be assumed to be a constant matrix. $\Omega(.)$ denotes a piece-wise linear regularization function that is used for implementing quantization operators. An example of a regularization function $\Omega(.)$ is shown in Fig. 3.4 for 1-dimensional input vector $\mathbf{v}$. Due to the piece-wise nature of the function

$\Omega(.)$ its gradient $\mathbf{d} = \nabla\Omega$ (shown in Fig. 3.4) is equivalent to scalar quantization operators. Without loss of generality, it will be assumed that the range of the quantization operator is limited between $[-1, 1]$. Therefore, for a $2K$ step quantization function the corresponding regularization function $\Omega(.)$ is given by

$$\Omega(\mathbf{v}) = \sum_{j=1}^{M} |\frac{i}{2K} v_j|; \quad |v_j| \in [i-1, i] \tag{3.6}$$

for $i = 1, .., 2K$.

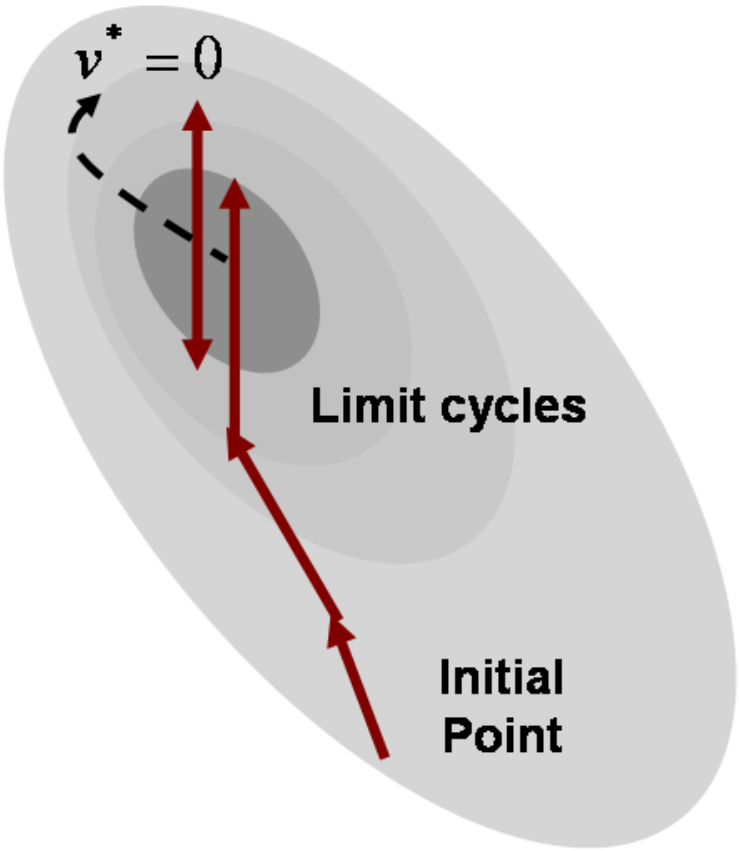



Figure 3.5: Limit cycle behavior using bounded gradients

To reiterate, the uniqueness of the proposed approach, compared to other optimization techniques to solve (3.4) is the use of bounded gradients to generate $\Sigma\Delta$ limit-cycles. This is illustrated in Fig. 3.5 showing the proposed optimization procedure using a two-dimensional contour.

Provided the input $\mathbf{x}$ and the norm of the linear transformation $||\mathbf{B}||_\infty$ are bounded and the regularization function $\Omega$ satisfies the Lipschitz condition, the optimal solution to (3.4) is well defined and bounded from below which is shown in the next lemma:

**Lemma 3.1.1.** *For the bounded matrix $||\mathbf{B}||_\infty \leq \lambda^{-1}$, bounded vector $||\mathbf{x}||_\infty \leq 1$, $\mathcal{C}$ as defined in equation (3.5) is convex and is bounded by below according to $\mathcal{C}^* = \min_{\mathbf{v}} \mathcal{C}(\mathbf{v}) > \frac{1}{2}(\frac{1}{K} - 1)$.*

*Proof.* A topological property of norms [96] will be used in this proof which states that for two integers $p, q$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$, the following relationship is valid for vectors $\mathbf{v}$ and $\mathbf{u}$

$$|\mathbf{v}^T \mathbf{u}| \leq ||\mathbf{v}||_p ||\mathbf{u}||_q \tag{3.7}$$

Setting $\mathbf{u} = \mathcal{E}_{\mathbf{X}}\{\mathbf{B}\mathbf{x}\}$ and applying equation (3.7) the following inequality is obtained:

$$||\mathbf{v}||_1 ||\mathbf{u}||_\infty \geq |\mathbf{v}^T \mathbf{u}| \geq \mathbf{v}^T \mathbf{u} \geq \mathbf{v}^T \mathcal{E}_{\mathbf{X}}\{\mathbf{B}\mathbf{x}\} \tag{3.8}$$

It can be easily verified that $\Omega(\mathbf{v}) \geq ||\mathbf{v}||_1 - \frac{1}{2}(\frac{1}{K} - 1)$ which is shown graphically in Fig. 3.4 for a one dimensional case and hence can be extended element-wise to the multi-dimensional case.

Using the definition of the matrix norm and the given constraints, it can be easily seen $||\mathbf{B}||_\infty \geq ||\mathbf{B}\mathbf{x}||_\infty \geq ||\mathbf{u}||_\infty$. Thus, $|\mathbf{u}||_\infty \leq \lambda^{-1}$. Therefore, the inequality (3.8) leads to

$$\Omega(\lambda^{-1}\mathbf{v}) - \mathbf{v}^T \mathcal{E}_{\mathbf{X}}\{\mathbf{B}\mathbf{x}\} \geq \lambda^{-1}||\mathbf{v}||_1 - \frac{1}{2}(\frac{1}{K} - 1) - \mathbf{v}^T \mathcal{E}_{\mathbf{X}}\{\mathbf{B}\mathbf{x}\} \geq 0 \tag{3.9}$$

which proves that the cost function $\mathcal{C}(\mathbf{v})$ is bounded from below by $\mathcal{C}^*$.

□

□

However, for $\Sigma\Delta$ learning the trajectory toward the minima of the cost function (3.5) is of importance. A stochastic gradient minimization corresponding to the optimization problem (3.5) leads to

$$\mathbf{v}[n] = \mathbf{v}[n-1] + \mathbf{B}\mathbf{x}[n] - \lambda^{-1}\mathbf{d}[n] \tag{3.10}$$

with $n$ signifying the time steps and $\mathbf{d}[n] = \nabla\Omega(\mathbf{v}[n-1])$ being the quantized representation according to functions shown in Fig. 3.4. Note also that formulation (3.10) does not require any learning rate parameters. As the recursion (3.10) progresses, bounded limit cycles are produced about the solution $\mathbf{v}^*$ (see Fig. 3.5) .

The following two lemma exploits the property of the first-order modulator to show that the auxiliary state variable $\mathbf{v}[n]$ defined by (3.10) is uniformly bounded if the input random vector $\mathbf{x}$ and matrix $\mathbf{B}$ are uniformly bounded.

**Lemma 3.1.2.** *For any bounded input vector sequence satisfying* $||\mathbf{x}[n]||_\infty \leq 1$ *and the transformation matrix* $\mathbf{B}$ *satisfying* $||\mathbf{B}||_\infty \leq \lambda^{-1}$*, the internal state vector* $\mathbf{v}[n]$ *defined by equation (3.10) is always bounded, i.e.,* $||\mathbf{v}[n]||_\infty \leq 2\lambda^{-1}$ *for* $n = 1, 2, ...$

*Proof.* The mathematical induction will be applied to prove this lemma. Without any loss of generality one can assume $||\mathbf{v}[0]||_\infty \leq 2\lambda^{-1}$. Suppose $||\mathbf{v}[n-1]||_\infty \leq 2\lambda^{-1}$, it then follows that $||\mathbf{v}[n-1] - \nabla\Omega(\mathbf{v}[n-1])||_\infty = ||\mathbf{v}[n-1] - \mathbf{d}[n]||_\infty \leq \lambda^{-1}$. Because $\mathbf{x}$ and $\mathbf{B}$ are bounded and using equation (3.10), the following relationship holds

$$\begin{aligned} ||\mathbf{v}[n]||_\infty &= ||\mathbf{v}[n-1] - \lambda^{-1}\mathbf{d}[n] + \mathbf{B}\mathbf{x}[n]||_\infty \\ &\leq ||\mathbf{v}[n-1] - \lambda^{-1}\mathbf{d}[n]||_\infty + ||\mathbf{B}||_\infty \\ &\leq 2\lambda^{-1} \end{aligned} \tag{3.11}$$

□

□

**Lemma 3.1.3.** *For any bounded input vector* $||\mathbf{x}||_\infty \leq 1$ *and bounded transformation matrix* $\mathbf{B}$*,* $\mathbf{d}[n]$ *asymptotically satisfies estimates* $\mathcal{E}_n\{\mathbf{d}[n]\} \overset{n\to\infty}{\longrightarrow} \lambda\mathcal{E}_n\{\mathbf{Bx}[n]\}$.

*Proof.* Following $N$ update steps the recursion given by equation (3.10) yields

$$\mathbf{Bx}[n] - \lambda^{-1}\mathcal{E}_n\{\mathbf{d}[n]\} = \frac{1}{N}(\mathbf{v}[N] - \mathbf{v}[0]) \tag{3.12}$$

which using the bounded property of random vector $\mathbf{v}$ asymptotically leads to

$$\mathcal{E}_n\{\mathbf{d}[n]\} \overset{n\to\infty}{\longrightarrow} \lambda\mathcal{E}_n\{\mathbf{Bx}[n]\} \tag{3.13}$$

□

□

Thus, according to Lemma 3, recursion (3.10) produces a quantized sequence whose mean asymptotically encodes the scaled transformed input at infinite resolution. It can also be shown that for a finite $I$ iterations of (3.10) yields a quantized representation that is $log_2(I)$ bits accurate.

### 3.1.1 $\Sigma\Delta$ Learning

In this section, the optimization framework will be extended to include on-line estimation of the transform $\mathbf{B}$. Here again $\lambda$ is assumed to be constant. Given an $M$ dimensional random input vector $\mathbf{x} \in \mathcal{R}^M$ and an internal state vector $\mathbf{v}$, the $\Sigma\Delta$ learning algorithm estimates parameters of a linear transformation matrix $\mathbf{B} \in \mathcal{R}^M \times \mathcal{R}^M$ according to the following optimization function

$$\max_{\mathbf{B}\in\mathcal{C}}(\min_{\mathbf{v}}\mathcal{C}(\mathbf{v},\mathbf{B})) \tag{3.14}$$

where

$$\mathcal{C}(\mathbf{v},\mathbf{B}) = \Omega(\lambda^{-1}\mathbf{v}) - \mathbf{v}^T\mathcal{E}_{\mathbf{X}}\{\mathbf{B}\mathbf{x}\}. \tag{3.15}$$

$\mathcal{C}$ denotes a constraint space on the transformation matrix $\mathbf{B}$. The minimization step in equation (3.14) will ensure that the state vector $\mathbf{v}$ is correlated with the transformed input signal $\mathbf{B}\mathbf{x}$ (tracking step) and the maximization step in (3.14) will adapt the matrix $\mathbf{B}$ such that it minimizes the correlation (de-correlation step).

The stochastic gradient descent step corresponding to the minimization yields the recursion

$$\mathbf{v}[n] = \mathbf{v}[n-1] + \mathbf{B}[n]\mathbf{x}[n] - \lambda^{-1}\mathbf{d}[n]. \tag{3.16}$$

where $\mathbf{B}[n]$ denotes the transform matrix obtained at time instant $n$. The transform $\mathbf{B}$ is then updated according to a stochastic gradient ascent step given by

$$\mathbf{B}[n] = \mathbf{B}[n-1] - 2^{-P}\mathbf{v}[n-1]\mathbf{x}[n]^T; \quad \mathbf{B}[n] \in \mathcal{C}. \tag{3.17}$$

$P$ in equation (3.17) is a parameter which determines the resolution of updates the parameter matrix $\mathbf{B}$. If we assume that locally the matrix $\mathbf{B}^*$ behaves as a positive definite matrix, equation (3.17) can be rewritten as

$$\begin{aligned}\mathbf{B}[n] &= \mathbf{B}[n-1] - 2^{-P}\mathbf{v}[n-1](\mathbf{B}[n]\mathbf{x}[n])^T \\ &= \mathbf{B}[n-1] - 2^{-P}\mathbf{d}[n]\mathbf{d}[n]^T\end{aligned} \tag{3.18}$$

where we have replaced the transformed input $\mathbf{B}[n]\mathbf{x}[n]$ by its asymptotic quantized representation $\mathbf{d}[n]$. Similarly $\mathbf{v}[n-1]$ is replaced by its quantized representation $\mathbf{d}[n]$. The update can be generalized further by incorporating non-linear quantization function $\phi(.)$ as

$$\mathbf{B}[n] = \mathbf{B}[n-1] - 2^{-P}\phi(\mathbf{d}[n])\mathbf{d}[n]^T \tag{3.19}$$

where $\phi : \mathcal{R}^M \rightarrow \mathcal{R}^M$ are functions dependent on the transformation $\mathbf{B}$. Here, $\phi(.) = \tanh(.)$ is assumed and the constraint space $\mathcal{C}$ has been chosen to restrict $\mathbf{B}$ to be a lower triangular matrix with all diagonal elements to be unity. One of the ways to ensure that $\mathbf{B}[n] \in \mathcal{C} \quad \forall n$ is to apply the updates only to lower triangular elements $b_{ij}; i > j$. The choice of this constraint guarantees convergence of the $\Sigma\Delta$ learning by ensuring $\mathbf{B}$ is bounded.

**Lemma 3.1.4.** *If the transform matrix* $\mathbf{B}$ *is bounded then the quantized sequences* $d_i[n]$ *and* $d_j[n]$ *with* $i \neq j$ *are uncorrelated with respect to each other.*

*Proof.* Using equation (3.19) the following relationships are obtained:

$$\begin{aligned} -2^{-P}\mathbf{d}[n]\phi(\mathbf{d}[n])^T &= \mathbf{B}[n] - \mathbf{B}[n-1] \\ -2^{-P}\mathcal{E}_n\{\mathbf{d}[n]\phi(\mathbf{d}[n])^T\} &= \lim_{N\rightarrow\infty}\frac{\mathbf{B}[N]}{N} \\ \mathcal{E}_n\{d_i[n]\phi(d_j[n])\} &= 0 \qquad \forall i \neq j \end{aligned} \tag{3.20}$$

Since this relationship holds for a generic form of $\phi(.)$, the sequences $d_i[n]$ are (non-linearly) uncorrelated with respect to each other. □

□

Equation (3.20) also provides a mechanism of reconstructing the input signal using the trans-

formed output $\mathbf{d}[n]$ and the converged estimate of the transformation matrix $\mathbf{B}[n] \overset{n\to\infty}{\longrightarrow} \mathbf{B}_\infty$. The input signal can be reconstructed using

$$\hat{\mathbf{x}} = \mathbf{B}_\infty^{-1}\lambda^{-1}\mathcal{E}_n\{\mathbf{d}[n]\}. \tag{3.21}$$

The use of lower-triangular transforms for $\mathbf{B}$ greatly simplifies the computation of the inverse $\mathbf{B}_\infty^{-1}$ through use of back-substitution techniques. Also, due to its lower-triangular form, the inverse of $\mathbf{B}_\infty$ always exists and is well defined.

### 3.1.2 Resolution Enhancement

Once the transform $\mathbf{B}$ has been determined such that the output of the $\Sigma\Delta$ learner is "de-correlated", we can apply resolution enhancement by "zooming" into the transformed signal space that does not cover the quantization regions (see Fig. 3.2(b)). This can be achieved using another diagonal matrix $\Lambda^{-1}$ which scales each axes as shown in the illustration 3.2(c). The $\Sigma\Delta$ cost function can be appropriately transformed to include the diagonal matrix $\Lambda \in \mathcal{R}^M \times \mathcal{R}^M$ as

$$C(\mathbf{v}, \mathbf{B}, \Lambda) = \Omega(\Lambda^{-1}\mathbf{v}) - \mathbf{v}^T\mathcal{E}_{\mathbf{X}}\{\mathbf{B}\mathbf{x}\}. \tag{3.22}$$

where the optimization (3.15) is also performed with respect to the parameter matrix $\Lambda$ such that the constraint $||B||_\infty < ||\Lambda^{-1}||_\infty$ is satisfied. This constraint is to ensure that $C(\mathbf{v}, \mathbf{B}, \Lambda)$ is always bounded from below. The stochastic gradient step equivalent to recursion (3.16) is given by

$$\mathbf{v}[n] = \mathbf{v}[n-1] + (\mathbf{B}[n-1]\mathbf{x}[n] - \Lambda^{-1}[n]\mathbf{d}[n]) \tag{3.23}$$

The asymptotic behavior of update (3.23) for equation (3.22) can be expressed as $\mathcal{E}_n\{d[n]\} \overset{n\to\infty}{\to} \Lambda\mathcal{E}_n\{\mathbf{B}_n\mathbf{x}_n\}$. Thus, reducing the magnitude of diagonal elements of $\Lambda$ will result in an equivalent amplification of the transformed signal. To satisfy the constraint on the transform $\mathbf{B}$ and $\Lambda$, a suitable update for the diagonal matrix $\Lambda$ and its elements $\lambda_i$ are

$$\lambda_i = \max|(\mathbf{B}[n]\mathbf{x}[n])_i|; N_1 > n > N_0 \tag{3.24}$$

where $N_0$ is the number of iterations required for the matrix $\mathbf{B}$ to stabilize and $N_1$ is the maximum observation period used to determine the parameters $\lambda_i$.

## 3.2 Acoustic source separation

Advances in acoustic miniaturization are enabling integration of an ever increasing number of microphones within a single sensor device which makes integration of miniature/compact microphone arrays possible. By introducing these devices, there have been several attempts in overcoming the fundamental problems introduced by these devices. A key challenge is to be able to image acoustical events occurring in the environment with high fidelity (spatial and temporal). However, due to the dispersive nature of the surrounding media, each element of the sensor array records a mixture of signals and noises generated by independent events in its environment. In order to improve the accuracy of the acoustic recognition systems, the signals of interest should be separated from the noises. The recovery of signals of interest from the recorded mixtures lies within the domain of blind source separation. Blind source separation (BSS) is based on a general class of unsupervised learning which has application in many areas of technologies. BSS task has connection to human perception where human hearing system has the ability to focus on acoustic sources of in-

terest even in a very noisy environment. Environmental assumptions about the surrounding of the microphone array directly influence the complexity of the BSS problem. Blind separation of the acoustic signals is sometimes referred to as the *Cocktail Party Problem* [90, 91] where the problem defined as the separation of voices from a mixture of sources in an uncontrolled environment like cocktail party. In the real world scenario, each microphone observation is a mixture of all the acoustic sources in the natural environment in which each of those acoustic sources are affected by signal reverberation. In order to make the problem more tractable, BSS techniques usually make some assumptions about the environment. The simplest scenario is termed *instantaneous mixing* where acoustic sources receive instantly at the microphones and only considering the intensity of sources. An extension of the previous assumption where arrival delays between microphones are also considered is know as the anechoic case. A more realistic assumptions lead to the *convolutive mixing* which considers multiple paths between each acoustic source and each microphone in order to model the signal reverberation. In modeling the BSS problem, assumptions can also be made about the number and statistical properties of the acoustic sources. It is very common to assume that sources are independent or at least decorrelated where the solution can be based on the higher order statistics (HOS) or second order statistics (SOS). This class of approaches are commonly called independent component analysis (ICA). A series of techniques are motivated by the insight from the auditory systems and they make strong assumption on acoustic sources such as common onset, harmonic structure, etc. These techniques are commonly refer to as computational auditory scene analysis (CASA) where they first detect and classify acoustic sources and then perform a supervised decomposition of the auditory scene. One increasingly popular and powerful assumption is that the acoustic sources have a sparse representation in some basis. These methods have come to be known as sparse methods. The advantage of a sparse signal assumption is that the probabil-

ity of two or more sources being active simultaneously is low. This results in good separability because most of the energy of the observed signal at any time instant belongs to a single source. It has also been shown that sparse representation exists in auditory cortex of brain in which firing pattern of neurons is characterized by long periods of inactivity [92, 93]. Usually it is assumed that there are at least as many sources as sensors for separation, but under strong assumption of sparsity it is sometimes possible to relax the conditions on the number of sensors. Some speech signal properties that can provide assumptions for BSS systems are as follows:

- Speech signals originating from different speakers at different spatial locations in an acoustic environment can be considered to be statistically independent.

- Speech signals are inherently non-stationary over long periods, but can be considered as quasi-stationary for small time durations (around 25 ms).

Most of linear BSS models can be expressed in the matrix format as:

$$\mathbf{X} = \mathbf{AS} + \mathbf{V} \tag{3.25}$$

where $\mathbf{X}$ is the observation matrix $\mathbf{X} \in \mathcal{R}^{m \times N}$, $m$ and $N$ being the number of observation and number of samples in each observation, $\mathbf{A} \in \mathcal{R}^{m \times N}$ represents the mixing matrix, $\mathbf{S} \in \mathcal{R}^{n \times N}$ contains the sources matrix, and $\mathbf{X} \in \mathcal{R}^{m \times N}$ is the noise matrix. Often BSS is performed by finding an $n \times m$, full rank separation matrix $\mathbf{W} = \mathbf{A}^{\dagger}$, where $\mathbf{A}^{\dagger}$ is some well-defined pseudo-inverse of $\mathbf{A}$ in a way that the output signal $\mathbf{Y} = \mathbf{WX}$ contains components that have special properties of interest based on the assumptions which can be measured by Kulllback-Leibler divergence or other criteria like sparseness, smoothness, etc.

## 3.3 Experimental results

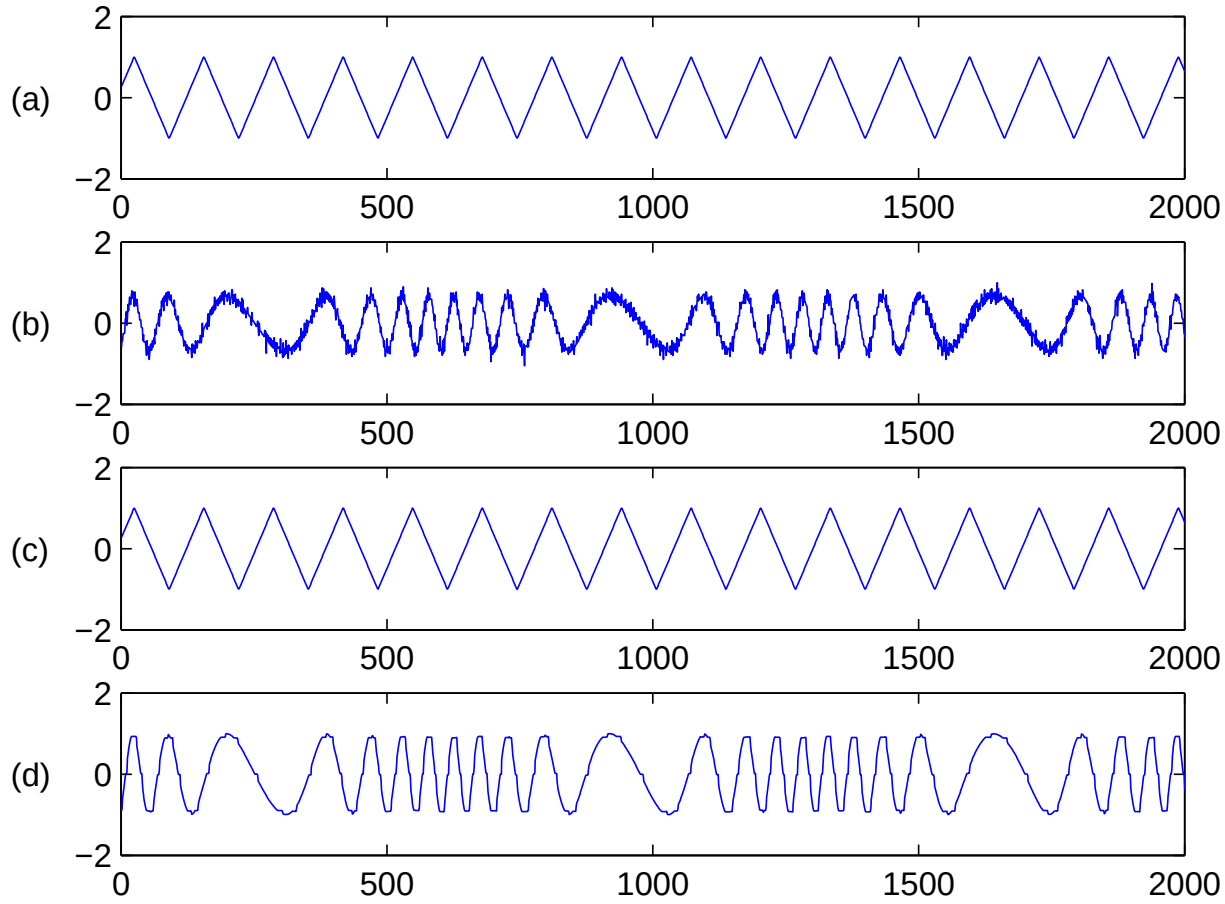



Figure 3.6: Reconstruction of the sources using conventional and proposed $\Sigma\Delta$ with OSR=1024

### 3.3.1 Numerical evaluation

The achievable improvement predicted by the equations (2.8) for $\Sigma\Delta$ learning was first verified using numerical evaluation. For this controlled experiment two synthetic signals were chosen.

$$\begin{aligned} \mathbf{s}_1(t) &= 480t - \lfloor 480t \rfloor \\ \mathbf{s}_2(t) &= \sin(800t + 6\cos(90t)) \end{aligned} \tag{3.26}$$

Each of these source signals were mixed using a random ill-conditioned matrix $\mathbf{A}$ to obtain the two-dimensional signals which were then processed by the $\Sigma\Delta$ learner. The outputs of the $\Sigma\Delta$ learner were then used to reconstruct the source signals according to

$$\tilde{\mathbf{s}}_d = \mathbf{A}^{-1}\mathbf{x} \tag{3.27}$$

$$\tilde{\mathbf{s}}_m = \mathbf{A}^{-1}\mathbf{B}^{-1}\mathbf{x} \tag{3.28}$$

$$\tilde{\mathbf{s}}_{\acute{m}} = \mathbf{A}^{-1}\mathbf{B}^{-1}\Lambda^{-1}\mathbf{x} \tag{3.29}$$

assuming that the un-mixing matrix $\mathbf{A}^{-1}$ can be perfectly determined.

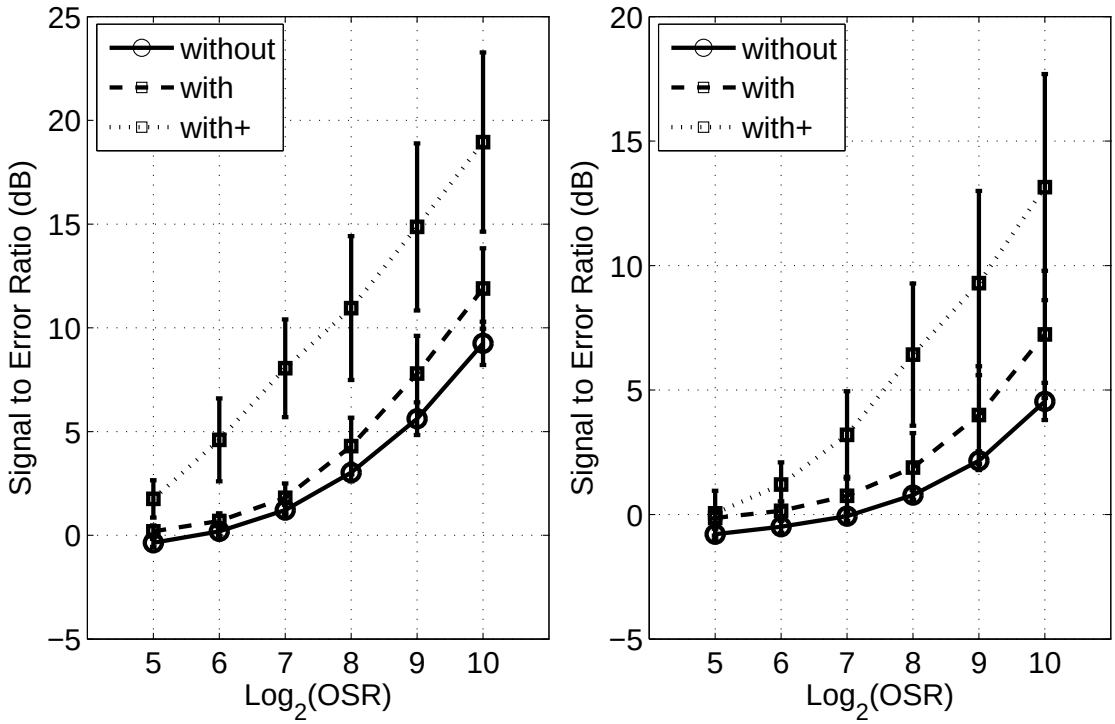



Figure 3.7: Evaluating the reconstruction of the sources for classical (without), learning (with), and learning with resolution enhancement (with+) $\Sigma\Delta$ at different OSR for $\log_2$(condition number) of (a) 10 and (b) 12

The equations (3.27)- (3.29) represent the following three cases: (a) $\tilde{\mathbf{s}}_d$ which is the signal

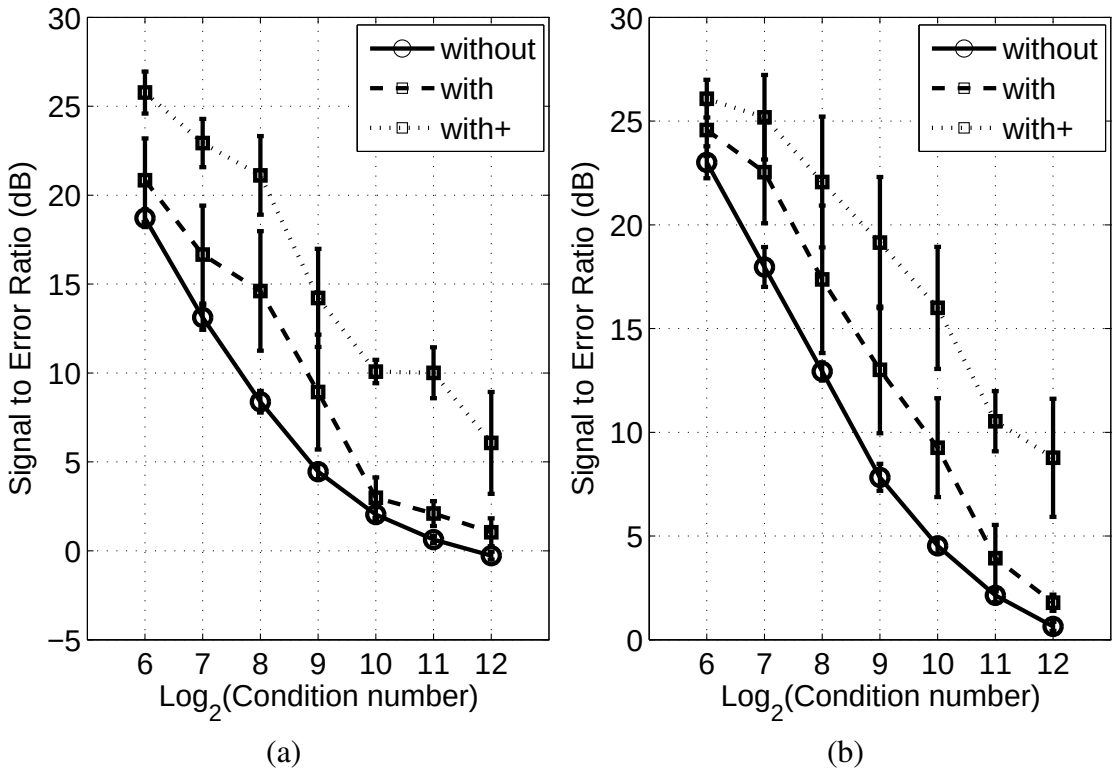



Figure 3.8: Evaluating the reconstruction of the sources for classical (without), learning (with), and learning with resolution enhancement (with+) $\Sigma\Delta$ at different condition number for OSR of (a) 256 and (b) 512

reconstructed using a $\Sigma\Delta$ modulator without any learning (denoted by *without*); (b) $\tilde{\mathrm{s}}_m$ which is the signal reconstructed using $\Sigma\Delta$ learning without resolution enhancement (denoted by *with*); and (c) $\tilde{\mathrm{s}}_{\acute{m}}$ which is the signal reconstructed using $\Sigma\Delta$ learning with resolution enhancement (denoted by *with+*). For this experiment, the condition number of the mixing matrix was chosen to be $1000$ and the over-sampling ratio (OSR), which is defined as the sampling frequency/Nyquist frequency, was chosen to be $1024$. For the signal in (3.26) the Nyquist frequency was chosen to be $10$ KHz. Figure 3.6 shows the reconstructed signals obtained with and without the application of $\Sigma\Delta$ learning. The quantization artifacts can be clearly seen Fig. 3.6(b) which is the signal recovered using $\Sigma\Delta$ modulator without learning. However, the signals obtained when $\Sigma\Delta$ learning is applied does not show any such artifacts indicating improvement in resolution. To quantify

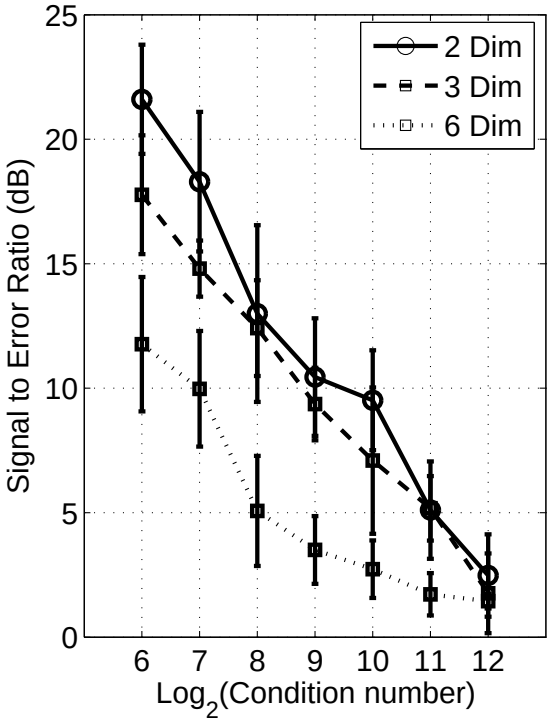



Figure 3.9: Evaluating the reconstruction of sources at different dimension for the learning $\Sigma\Delta$ at different condition number for OSR of 128

this improvement, we compared the signal-to-error ratio (SER) for the separated signals. SER is defined as

$$SER = \log_2\{\frac{||\mathbf{s}||_2}{||\mathbf{s} - \tilde{\mathbf{s}}||_2}\} \tag{3.30}$$

where $\mathbf{s}$ and $\tilde{\mathbf{s}}$ is based on the definition in (3.27)- (3.29). To compute the mean SER and its variance $10$ different mixing matrices with a fixed condition number were chosen and the mean/variances were calculated across different experimental runs. Fig. 3.7(a) compares the SER obtained when the mixing matrix with condition number $2^{10}$ was chosen for different values of OSR. Figure 3.7(b) compares the SER obtained when the mixing matrix with condition number $2^{12}$ was chosen. It can be seen in Fig. 3.7(a) and (b) that as the OSR of the $\Sigma\Delta$ modulator increases, the SER increases. This is consistent with results reported for $\Sigma\Delta$ modulators where OSR is directly related to the resolution of the "analog-to-digital" conversion. However, it can be seen that for all conditions of

OSR, $\Sigma\Delta$ learning with resolution enhancement outperforms the other two approaches.

Figure 3.8(a) and (b) compares the performance of $\Sigma\Delta$ learner when the condition number of the mixing matrix is varied for fixed over-sampling ratios of $256$ and $512$. The results again show that $\Sigma\Delta$ learner (with and without resolution enhancement) demonstrates consistent performance improvement over the traditional $\Sigma\Delta$ modulator. Also, as expected the SER performance for all the three cases deteriorates with the increasing condition number, which indicates that the mixing becomes more singular. Figure 3.9 evaluates the SER achieved by $\Sigma\Delta$ learning (with resolution enhancement) when the dimensionality of the mixing matrix is increased. For this experiment, the number of source signals are increased by randomly selecting signals which were mutually independent with respect to each other. It can be seen from Fig. 3.9 that the response of the $\Sigma\Delta$ learning is consistent across different signal dimensions with larger SER when the dimension is lower.

In the next set of experiments the performance of $\Sigma\Delta$ learning is evaluated for the task of source separation when the un-mixing matrix is estimated using an ICA algorithm. Speech samples were obtained from TIMIT database and were synthetically mixed using an ill-conditioned matrix with different condition number. The instantaneous mixing parameters simulate the "near-far" scenario where one of the speech sources is assumed to be much closer to the microphone array than the other. This scenario was emulated by scaling one of the signals by $-50dB$ with respect to another. The speech mixture is then presented to the $\Sigma\Delta$ learner and its output is then processed by a second-order blind inference (SOBI) [97] and by an efficient FastICA (EFICA) [98] algorithms. The performance metrics chosen for this experiment is based on source-to-distortion ratio (SDR) [99] where the estimated signal $\hat{s}_j(n)$ is decomposed into

$$\hat{s}_j(n) = s_{target}(n) + e_{interf}(n) + e_{artif}(n) \tag{3.31}$$

with $s_{target}(n)$ being the original signal, and $e_{interf}(n)$ and $e_{artif}(n)$ denote the interference and artifacts errors, respectively. The SDR metric is a global performance metric which then measures both source-to-interference ratio (SIR) - the amount of interferences from non wanted sources and also other artifacts like quantization and musical noise. The SDR is defined as:

$$SDR = 10\log\frac{\|s_{target}\|^2}{\|e_{interf}+e_{artif}\|^2} \tag{3.32}$$

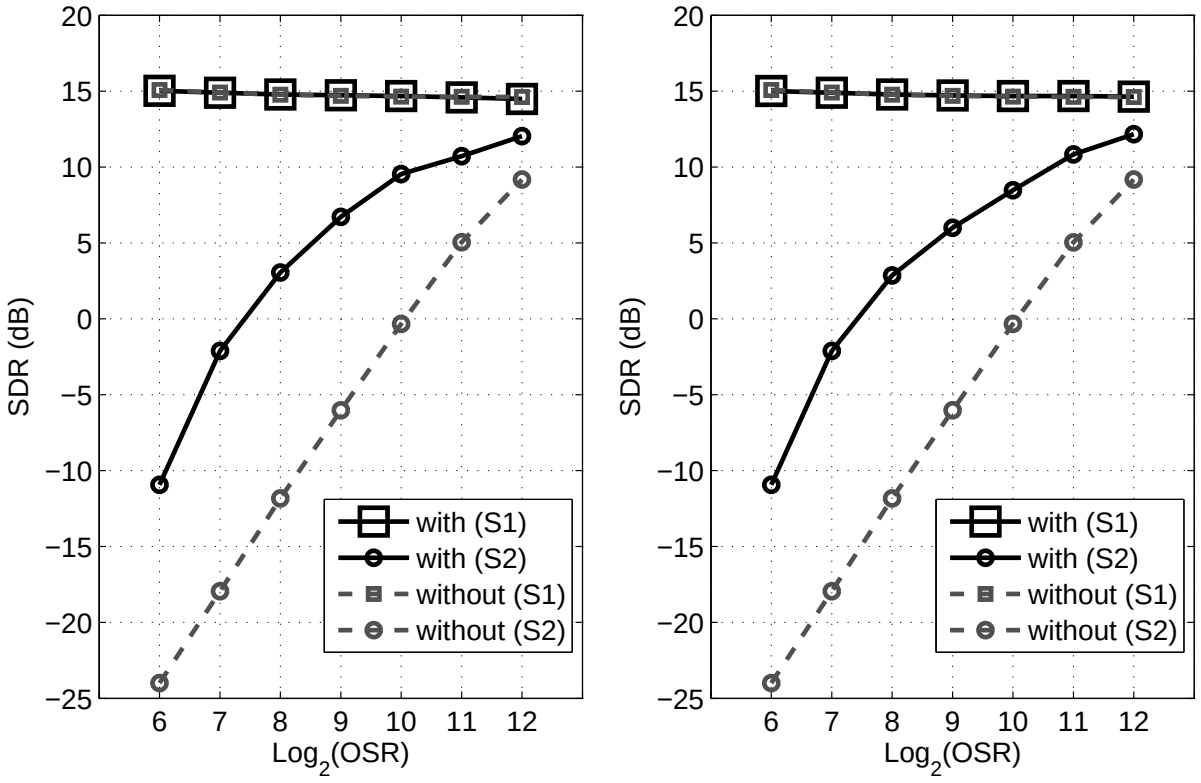



Figure 3.10: SDR corresponding to with/without $\Sigma\Delta$ learning for the near-far recording conditions using (a) SOBI and (b) EFICA algorithms.

The speech sources $s_1$ and $s_2$ in this experiment consists of 44200 samples with a sampling rate of 16KHz which is also the Nyquist rate. In this setup, after mixing, one of the sources is being completely masked by the other which is consistent with the"near-far" effect. Figure 3.10(a)

and (b) shows the SDR obtained using the $\Sigma\Delta$ learning when the OSR is varied from 128 to 4096 . Also shown in Fig. 3.10 (a) and (b) are the SDR metrics obtained when a conventional $\Sigma\Delta$ algorithm is used. It can be seen from the Fig. 3.10 that the SDR corresponding to the stronger source is similar for both cases (with and without $\Sigma\Delta$ learning), where as for the masked source the SDR obtained using $\Sigma\Delta$ learning is superior. This is consistent with the results published in [100]. However, the approach in [100] is applied after quantization and hence according to formulation in section I, is limited by the condition number of the mixing matrix. It should also be noted that the $\Sigma\Delta$ learning only enhances the resolution of the measured signals. The ability to successfully recover the weak source under "near-far" conditions, however, is mainly determined by the choice of the ICA algorithm.

### 3.3.2 Experiments with far-field model

In this section, the mathematical model presented in chapter 2 for miniature microphone model will be used. This model with the $\Sigma\Delta$ learner will be used to compare the performance of the algorithm with a traditional source separation technique. Traditional DSP based source separation algorithms are typically implemented subsequent to a ADC and hence do not consider the detrimental effects of finite resolution due to the quantizer. $\Sigma\Delta$ learner smartly quantizes the array signals with respect to each other and uncorrelates them in order to use as much information as possible when digitizing the signals.

In this setup, recording conditions consisted of four closely spaced microphones. In this arrangement, three of the microphones were placed along a triangle, whereas the fourth microphone was placed at the centroid and act as the reference sensor which records the common signal. The set up is similar to the conditions that have been reported in [86] where the simulation have been shown

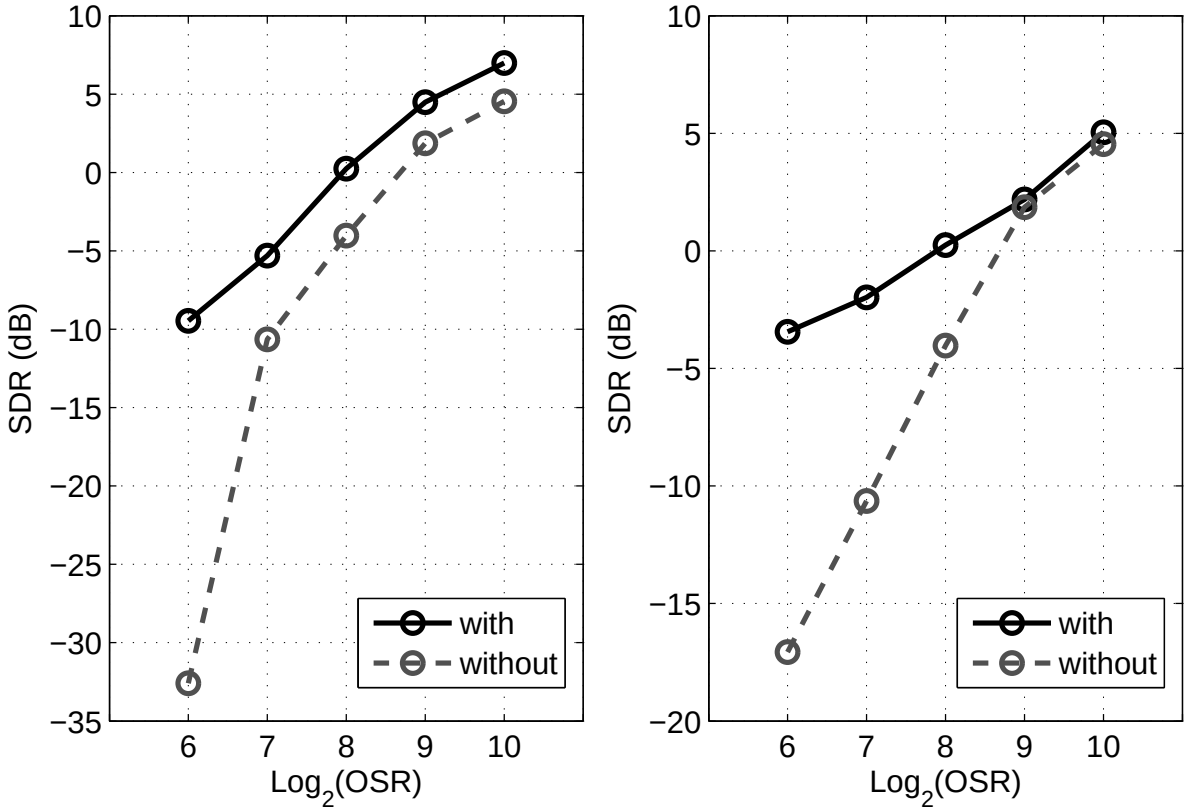



Figure 3.11: $\Sigma\Delta$ performance with and without learning for three speech signals corresponding to the far-field model

to be consistent with recording in the real-life conditions. The outputs of each microphone along the triangle were subtracted from the reference microphone to produce three differential outputs. In these experiments three independent speech signals were used as far-field sources. The differential outputs of the microphone array were first presented to the proposed $\Sigma\Delta$ learner, and the outputs of $\Sigma\Delta$ learner array were then used as inputs to the SOBI algorithm. A benchmark used for comparative study consisted of $\Sigma\Delta$ converters that directly quantized the differential outputs of the microphones. Figure 3.11(b)-(c) summarizes the performance of source separation (with and without $\Sigma\Delta$ learning) for different orientation of the acoustic sources. For the three experiments, only the bearing of the sources were varied but their respective distances to the center of the mi-

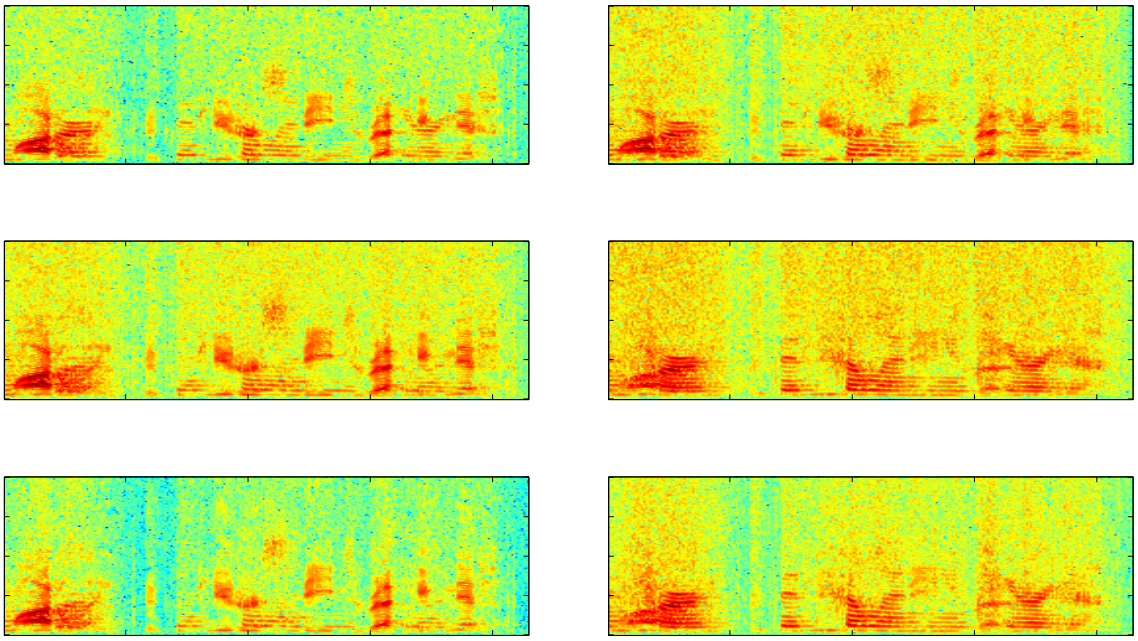


Figure 3.12: Spectrogram of the recorded signals (top row) and recovered signals using $\Sigma\Delta$ without learning (middle row) and with learning (bottom row)

crophone was kept constant. It can be seen that for each of these orientations, source separation algorithm that uses $\Sigma\Delta$ learning as a front-end "analog-to-digital" converter delivers superior performance compared to the algorithm that does not use $\Sigma\Delta$ learning. Also, from Fig. 3.11(b)-(c) it can be seen that the improvement in SDR performance significantly increases when the sampling frequency (resolution) decreases showing that $\Sigma\Delta$ learning efficiently utilizes the available resolution (due to coarse quantization) to capture information necessary for source separation.

### 3.3.3 Experiments with real microphone recordings

In this set of experiments, $\Sigma\Delta$ learning have been applied to speech data that was recorded using a prototype miniature microphone array, similar to the set up described in [85]. The four omnidirectional microphones *Knowles FG3629* were mounted on a circular array with radius 0.5 cm.

The differential microphone signals were pre-processed using second-order bandpass filters with low-frequency cutoff at 130 Hz and high-frequency cutoff at 4.3 kHz. The signals were also amplified by 26dB. The speech signals were presented through loudspeakers positioned at 1.5 m distance from the array and the sampling frequency of the National Instruments data acquisition card was set to 32 KHz. Male and female speakers from TIMIT database were chosen as sound sources and was replayed through standard computer speakers. The data was recorded from each of the microphones, archived, and then presented as inputs to the $\Sigma\Delta$ learning and the SOBI algorithm. Figure 3.12(top row)) shows the spectrogram of the speech signals recorded from the microphone array. The two spectrograms look similar, thus emulating a "near-far" recording scenario where a dominant source masks the background weak source. Also it can be seen from the spectrogram in Fig. 3.12(top row-right) that one of the recordings is more noisy than the other (due to microphone mismatch). Figure 3.12(middle row) show the spectrogram of the separated speech signals obtained without $\Sigma\Delta$ learning. Figure 3.12(bottom row) show the spectrogram of the separated speech signals obtained with $\Sigma\Delta$ learning. A visual comparison of the spectrograms show that separated speech signal (without $\Sigma\Delta$ learning) contain more quantization artifacts which can be seen as the broadband noise in Fig. 3.12(middle row). The table 3.1 summarizes the SDR performance (for different OSR) for each of the sources in these two cases (with and without $\Sigma\Delta$ learning), showing that $\Sigma\Delta$ learning indeed improves the performance of the source separation algorithm. Also, from table 3.1, it can be seen that when the OSR increases, the performance differences between the two cases becomes insignificant. This artifact is due to the limitations in the SOBI algorithm for separating sources with high fidelity, noise in the microphones and ambient recording conditions.

Up to here, It have been argued that the classical approach of signal quantization followed by

Table 3.1: Performance (SDR (dB) ) of the proposed $\Sigma\Delta$ for the real data for different over-sampling ratio.

| | OSR=4 | | OSR=8 | | OSR=16 | | OSR=32 | | OSR=64 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | with | without | with | without | with | without | with | without | with | without |
| $S_1$ | 1.03 | -0.72 | 1.33 | 0.41 | 1.34 | 0.94 | 1.3 | 1.15 | 1.28 | 1.21 |
| $S_2$ | -12.52 | -13.15 | -9.77 | -10.17 | -8.69 | -8.88 | -8.29 | -8.32 | -8.12 | -8.1 |

DSP based source separation fails to deliver robust performance when processing signals recorded using miniature microphone/sensor arrays. We proposed a framework that combined statistical learning with $\Sigma\Delta$ modulation and can be used for designing "smart" multi-dimensional analog-to-digital converters that can exploit spatial correlations to resolve acute differences between signals recorded by miniature microphone array.

# Chapter 4

# Robust Acoustic Recognition

## 4.1 Fundamental of speech

Speech is produced when air from the lungs passes through the throat, the vocal cords, the mouth and the nasal tract (see Fig. 4.1(a)). Different position of the lips, tongue and the palate (also known as the articulators) then create different sound patterns and gives rise to the physiological and spectral properties of the speech signal like pitch, tone and volume. These properties are speaker related and they can be used as signitures for speaker recognition systems as they are modulated by the size and shape of the mouth, vocal and nasal tract along with the size, shape and tension of the vocal cords of each speaker where It has been shown that even for twins, the chances for all of these properties to be similar are very low [32, 33].

One of the most commonly used methods for visualizing the spectral and dynamical content of speech signal is called the spectrogram which displays the frequency of vibration of the vocal cords (pitch), and amplitude (volume) with respect to time. Examples of the spectrograms for a male and a female speaker are shown in Fig. 4.1(b) where the horizontal axis represents time and the vertical

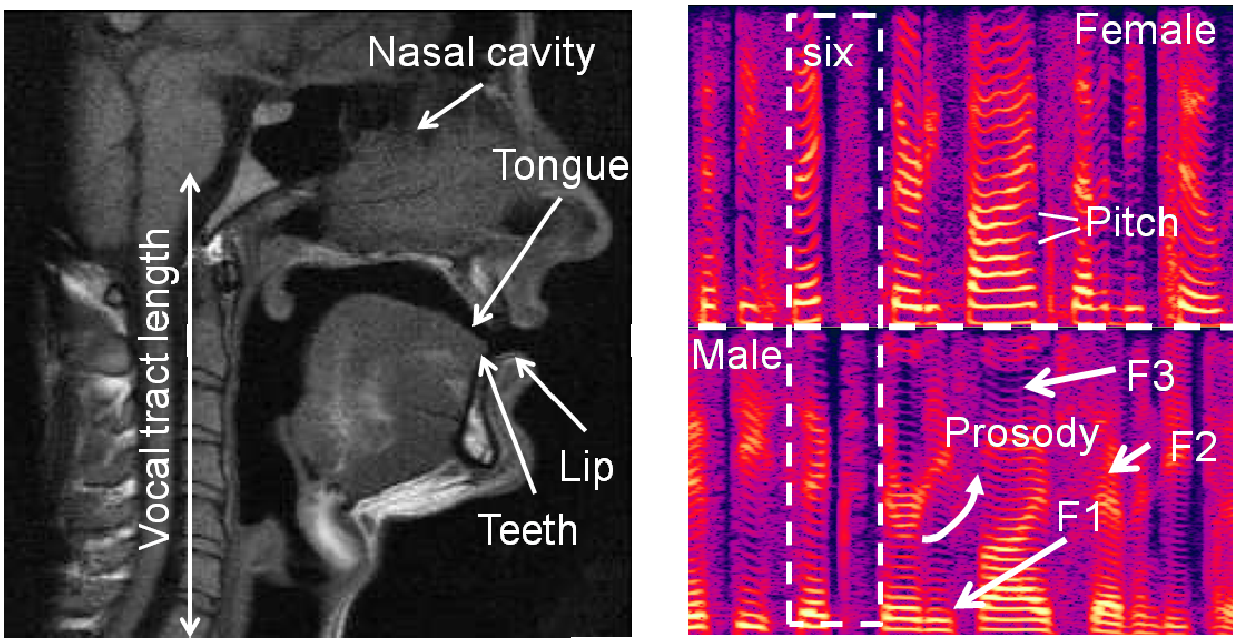



Figure 4.1: Fundamental of speech: (a) Magnetic resonance image showing the anatomy of speech production apparatus. The property of the speech signal is determined by shape of the vocal tract, orientation of the mouth, teeth and nasal passages. (b) Spectrograms corresponding to a sample utterance "fiftysix thirty-five seventy-two" for a male and female speake.

axis represents frequency. The pitch of the utterance manifests itself as horizontal striations in the spectrogram as shown in Fig. 4.1(b). For instance, it can be seen from Fig. 4.1(b) that the pitch of the female speaker is greater than the pitch of the male speaker. Other important spectral parameters of speech signal are formants which are defined as the resonant frequencies (denoted by F1, F2, F3, ...) of the vocal tract, in particular, when vowels are pronounced. They are produced by restricting air flow through the mouth, tongue, and the jaw. The relative frequency location of the formats can vary widely from person to person (due to shape of the vocal tracts) and hence can be used as a biometric feature. Even though multiple resonant frequencies exist in speech signal, only three of the formats (typically labeled as F1, F2, F3 as shown in Fig. 4.1(b)) are used for speech and speaker recognition applications. However, reliable estimation of the spectral parameters requires segments of speech signal that are stationary and hence most verification systems use 20-30

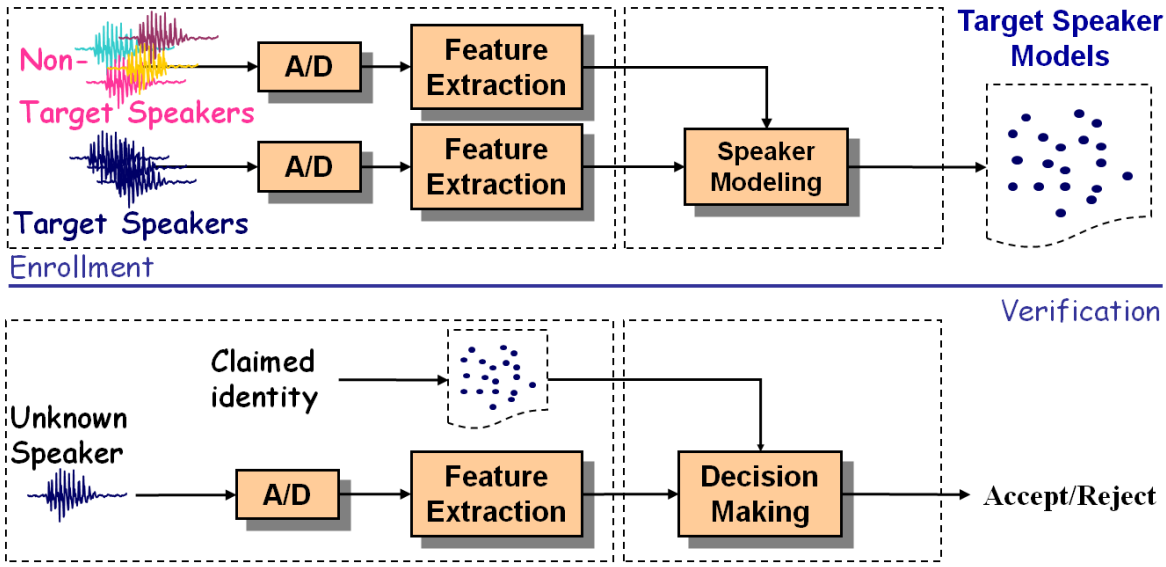



Figure 4.2: Functional architecture of a speaker verification system as a example of acoustic recognition which consists of two main phases: (a) An training/enrollment phase where parameters of a speaker specific statistical model are determined and (b) a recognition/verification phase where an unknown speaker authenticated using the models trained during the training phase.

milliseconds segments. Another biometric signature embedded in the speech signal is the stress patterns also known as prosody which manifests as spectral trajectory and distribution of energy in the spectrogram. This signature is typically considered as one of the "high-level" features which can be estimated from observing the dynamics across multiple segments of the speech signals. In the next section, we will discuss some of the popular approach to extract some of these biometric features and discuss some of the statistical models which are used to recognize the speaker specific features.

## 4.2 Architecture of an acoustic recognition system

Any speech based recognition systems like speaker or speech recognition typically consist of two distinct phases in general: (a) a training phase where parameters of statistical models are determined using annotated (pre-labeled) speech data; and (b) a testing phase where an unknown speech

sample is recognized using the trained statistical models. Fig. 4.2 presents an speaker verification system where these two phases are shown as enrollment and verification phases. As this figure shows, in such a system the speech signal is first sampled, digitized, and filtered before a feature extraction algorithm computes salient acoustic features from the speech signal. The next step in the training phase uses the extracted features to train a statistical model. During the recognition phase (as shown in Fig. 4.2), an unknown utterance is authenticated against the trained statistical model for a specific task. In the following sections, each of these standard modules will be reviewed that are used during each of these phases.

## 4.3 Speech acquisition and feature extraction module

The speech acquisition module typically consists of a transducer that is coupled to an amplifier and a filtering circuitry. Depending on the specifications (size, power and recognition performance) imposed on the recognition system, the transducer could be a standard microphone (omni-directional or directed) or a noise-canceling microphone array where the speech signal is enhanced by suppressing background noise using a spatial filter [34]. The amplifier and the filtering circuitry are used to maintain a reasonable signal-to-noise ratio (SNR) at the input of an analog-to-digital converter (ADC) which is used to digitize the speech signal. Depending on the topology of the ADC, the speech signal could be sampled at the Nyquist rate (8KHz) or oversampled using a sigma-delta modulator. Typically, a high-order sigma-delta modulator is the audio ADC of choice because of its ability to achieve resolution greater than 16 bits. Once the speech signal is digitized, a feature extraction module (typically implemented on a digital signal processor) extracts speech information from the raw waveform. Depending on the application of the recognition system, different feature can be extracted, *e.g.,* in speaker recognition system feature extraction module extracts

speaker specific features where in speech recognition systems, this module extracts type of features that are more speaker independent. In speaker recognition systems the "high-level" characteristics which convey behavioral information such as prosody, phonetic, conversational patterns, etc. seem to be promising than the "low-level" information which conveys the physical structure of the vocal tract [117, 103]. The "low-level" features have been mostly used in speech recognition systems, however it has been shown that good performances can be achieved using these features for speaker recognition systems. The difference between these two features is the relative time-scale required for extracting and processing the features. While "low-level" features can be effectively computed using short frames of speech ($<$30ms), the "high-level" features could require time-scales greater than few seconds [103]. In the following, we present a short overview of two of the popular classes of "low-level" features: linear predictive cepstral coefficients (LPCC) and Mel frequency cepstral coefficients (MFCC).

**Linear Predictive Cepstral Coefficients (LPCC):** The basic assumption underlying the Linear Prediction Coding (LPC) [101, 104], which is in the heart of LPCC is that speech signal can be modeled by a linear source-filter model. This model has two sources of human vocal sounds: the glottal pulse generator and the random noise generator. The glottal pulse generator creates voiced sounds. This source generates one of the measurable attributes used in voice analysis: the pitch period. The random noise generator produces the unvoiced sounds and the vocal tract serves as the filter of the model that produces intensification at specific formants. In LPC feature extraction, the filter is typically chosen to be an all-pole filter. The parameters of the all-pole filter are estimated using an auto-regressive procedure where the signal at each time instant can be determined using

a certain number of preceding samples. Mathematically, the process can be expressed as

$$s(t) = -\sum_{i=1}^{P} a_i s(t-i) + e(t)$$

where $s(t$ is the speech signal at time instant $t$ is determined by $p$ past samples $s(t-i)$ where $i$ represents the discrete time delay. $e(t)$ is known as the excitation term (random noise or glottal pulse generator) which also signifies the estimation error for the linear prediction process and $a_i$ denotes the LPC coefficients. During an LPC, a quasi-stationary window of speech (about 20-30ms) is used to determine the parameters $a_i$ and the process is repeated for the entire duration of the utterance. In most implementations, an overlapping window or a spectral shaping window [104] is chosen to compensate for spectral degradation due to finite window size. The estimation of the prediction coefficients is done by minimizing the prediction error $e(t)$ and several efficient algorithms like the Yule-Walker or Levinson Durbin algorithms exist to compute the features in real-time. The prediction coefficients will be further transformed into Linear Predictive Cepstral Coefficients (LPCC) using a recursive algorithm [104]. A variant of the LPC analysis is the Perceptual Linear Prediction (PLP) [60] method. The main idea of this technique is to take advantage of some characteristics derived from the psychoacoustic properties of the human ear and these characteristics are modeled by filter-bank.

**Mel Frequency Cepstral Coefficients (MFCC):** These features have been extensively used in speech based recognition systems [105, 104]. MFCCs were introduced in early 1980s for speech recognition applications and since then have also been adopted for speaker recognition applications. A sample of speech signal is first extracted using a window. Typically two parameters are important for the windowing procedure: the duration of the window (ranges from 20 - 30 ms) and the shift between two consecutive windows (ranges from 10-15ms). The values correspond to the

average duration for which the speech signal can be assumed to be stationary or its statistical and spectral information does not change significantly. The speech samples are then weighed by a suitable windowing function, for example, Hamming or Hanning window are extensively used in acoustic recognition. The weighing reduces the artifacts (side lobe and signal leakage) of choosing a finite duration window size for analysis. The magnitude spectrum of the speech sample is then computed using a fast Fourier transform (FFT) and is then processed by a bank of band-pass filters. The filters that are generally used in MFCC computation are triangular filters, and their center frequencies are chosen according to a logarithmic frequency scale, also known as Mel-frequency scale. The filter bank is then used to transform the frequency bins to Mel-scale bins by the following equations:

$$m_y[b] = \sum_f w_b[f]|Y[f]|^2$$

where $w_b[f]$ is the $b^{\text{th}}$ Mel-scale filter's weight for the frequency $f$ and $Y[f]$ is the FFT of the windowed speech signal. The rationale for choosing a logarithmic frequency scale conforms to response observed in human auditory systems which has been validated through several biophysical experiments [104]. The Mel-frequency weighted magnitude spectrum is processed by a compressive non-linearity (typically a logarithmic function) which also models the observed response in a human auditory system. The last step in MFCC computation is a discrete cosine transform (DCT) which is used to de-correlate the Mel-scale filter outputs. A subset of the DCT coefficients are chosen (typically the first and the last few coefficients are ignored) and represent the MFCC features used in the training and the test phases.

**Dynamic and Energy Features:** Even though each feature set (LPC or MFCC) is computed for a short frame of speech signal (about 20-30ms), it is well known that information embedded in the temporal dynamics of the features are also useful for recognition [106]. Typically two kinds

of dynamics have been found useful in speech processing: (a) velocity of the features (known as $\Delta$ features) which is determined by its average first-order temporal derivative; and (b) acceleration of the features (known as $\Delta\Delta$ features) which is determined by its average second-order temporal derivative. Other transforms of the features which have also been found useful in recognition include: logarithm of the total energy of the feature (L2 norm) and its first-order temporal derivative [104].

**Auxiliary Features:** Even though cepstral features have been widely used speaker recognition systems, it can be suggested that the features might contain phonemic information that may be unrelated to the speaker recognition task as they convey less speaker specific information. Recently, new techniques have been reported that can extract speaker-related information from LPCCs and MFCCs and in the process improve system's recognition performance. One group of these features are sometimes referred to as voice source features. For example, in [107], an inverse filtering technique has been used to separate the spectra of glottal source and vocal tract. In another approach, the residual signal obtained from LP analysis has been used in estimating the glottal flow waveform [108, 109, 110, 111]. An alternative approach to estimating the glottal flow (derivative) waveform was presented in [112, 113, 114] where a closed-phase covariance analysis technique was used during the intervals when the vocal folds are closed. Other group of these features includes prosodic features. Prosody which involves variation in syllable length, intonation, formant frequencies, pitch, rate and rhythm speech, can vary from speaker to speaker and relies on long-term information of speech. One of the predominant prosodic features is the fundamental frequency (or F0). Other features include, pitch, energy distribution on a longer frame, speaking rate and phone duration [115, 116, 117]. The auxiliary features usually have been used in addition to"low-level" features by fusion technique

**Voice Activity Detector (VAD):** Before the features can be used in the recognition systems it is important to determine whether the features correspond to the ”speech” portion of the signal or correspond to the silence or background part of the signal. Most speech based recognition systems use a voice activity detector (VAD) whose function is to locate the speech segments in an audio signal. For example, a simple VAD could compute instantaneous signal-to-noise ratio (SNR) and pick segments only when the SNR exceeds a predetermined threshold. However, it is improtant to know that design of robust VAD could prove challenging since it is expected that the module works consistently across different environments and noise conditions.

## 4.4 Speech and speaker modeling

Once the feature vectors corresponding to the speech frames have been extracted the associated speech data also known as training data is used to build models even for speech or speaker recognition systems. For speech recognition systems, the models are generated for the speech components like phonemes, words, etc. and for the speaker recognition systems, speaker models are generated. During the test phase, the trained model is used to recognize a sequence of feature vectors extracted from unknown utterances. The focus of this section is on the statistical approaches for constructing the relevant models. The methods can be divided into two distinct categories: *generative* and *discriminative*. Training of generative models typically involve data specific to the target speech component or speaker with the objective that the model can faithfully capture the statistics of that component. Training of discriminative models which have been used more in speaker recognition systems, involves data corresponding to the target and imposter speakers and the objective is to faithfully capture the manifold which distinguishes the features for the target speakers from the features for the imposter speakers. An example of a popular generative model used in speaker ver-

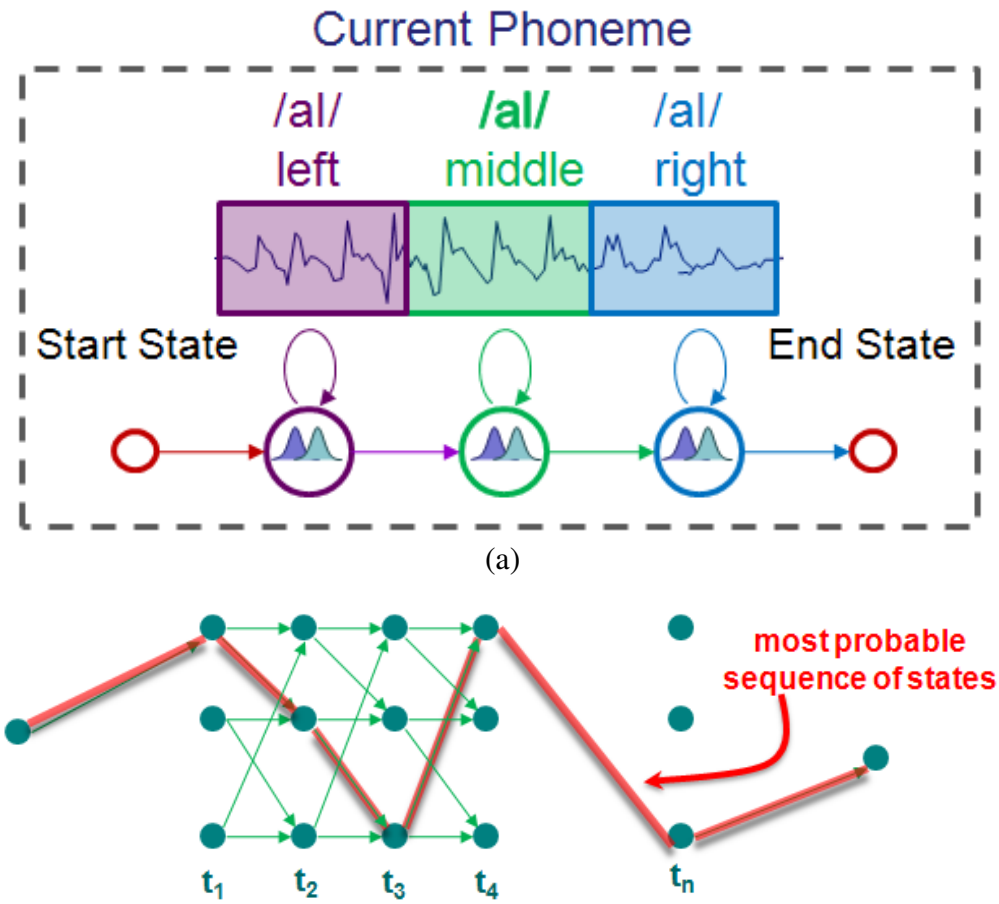



Figure 4.3: Example of generative models that have been used for speech/speaker recognition: (a) HMMs where each state has a GMM which captures the statistics of a stationary segment of speech. (b) HMMs are trained by aligning the states to the utterance using a trellis diagram. Each path through the trellis (from start to end) specifies a possible sequence of HMM state that generated the utterance

ification is *Gaussian Mixture Models (GMMs)*, and an example of a popular discriminative model is *Support Vector Machines (SVMs)*. In the following section these classical techniques will briefly be described. *Hidden Markov Models (HMMs)* is also a generative model which have been extensively used in speech recognition systems. In the following section these classical techniques will briefly be described and the readers are referred to appropriate references [104] for details.

### 4.4.1 Generative Models

Generative models include mainly Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) capture the empirical probability density function corresponding to the acoustic feature vectors. GMMs represent a particular case of HMMs and can be viewed as a single-state HMM where the probability density is defined by a mixture of Gaussians.

**GMM-based modeling.** GMMs have unique advantages compared to other modeling approaches because their training is relatively fast and the models can be scaled and updated to add new speakers with relative ease. A GMM model $\lambda$, is composed of a finite mixture of multivariate Gaussian components and estimates a general probability density function $p_{\lambda}(\mathbf{x})$ according to:

$$p(\mathbf{x}) = \sum_{i=1}^{M} w_i p_i(\mathbf{x})$$

where $M$ is the number of Gaussian components, $w_i$ is the prior probability (mixing weights) of the $i^{\text{th}}$ D-variate Gaussian density function $\mathcal{N}_i(\mathbf{x})$ given by

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}$$

The parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ represent the mean vector and covariance matrix of the multi-dimensional Gaussian distribution and the mixing weights $w_i$ are constrained according to $\sum_{i=1}^{M} w_i = 1$ .

GMM have extensively been used in speaker recognition system. Usually in these systems, a speaker-independent world model also known as a universal background model (UBM) is first trained using speech data gathered from a large number of *imposter* speakers [118]. The training procedure typically uses an iterative expectation-maximization (EM) algorithm [119] which estimates the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ using a maximum likelihood criterion [120]. More details on EM training procedure can be found in numerous references [119, 120, 121]. The background model obtained after the training thus represents a speaker-independent distribution of the feature vectors. When enrolling a new speaker to the system, the parameters of the background model are adapted to the feature vector distribution of the new speaker using the maximum a posteriori (MAP) update rules. In this way, the model parameters are not required to be estimated from scratch and instead the previously estimated priors are used for re-training. There are alternative adaptation methods to MAP, and usually selection of the method depends on the amount of available training data [122]. For very short enrollment utterances (a few seconds), some other methods like Maximum likelihood linear regression (MLLR) [52], have shown to be more effective.

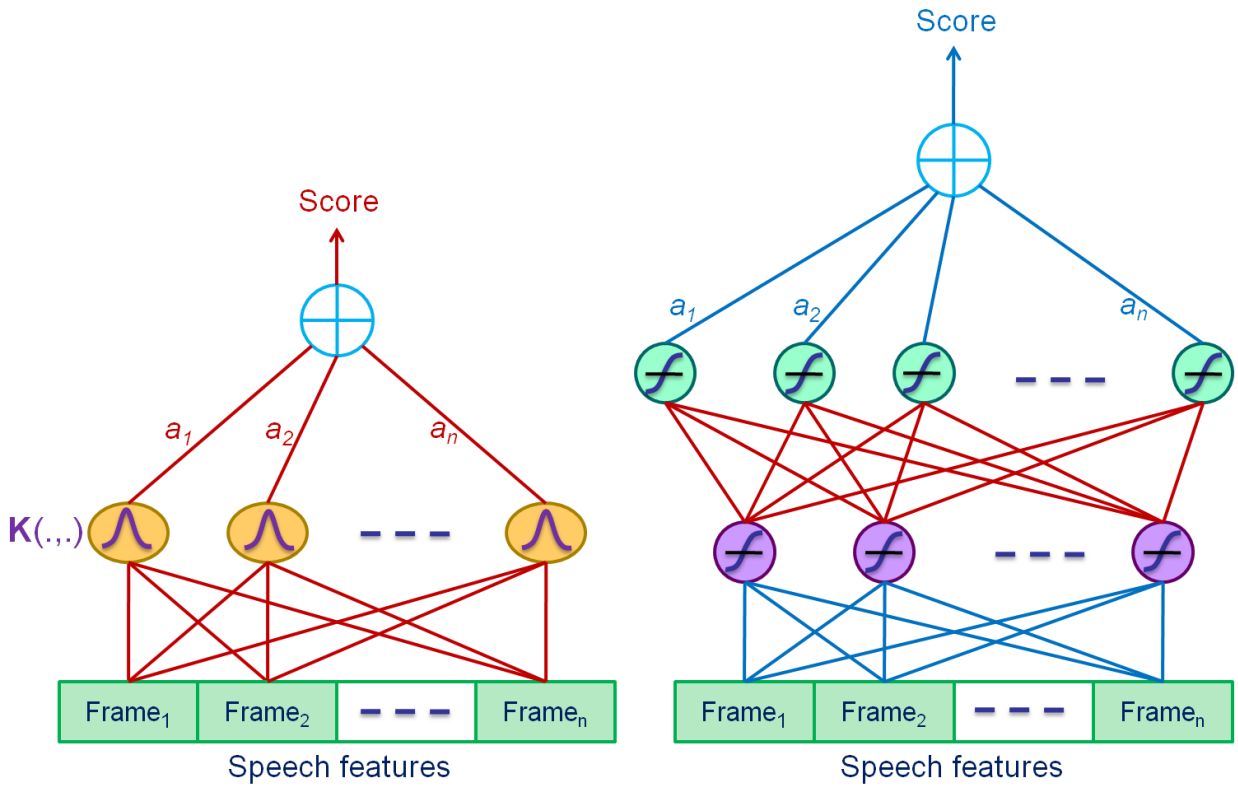
**Hidden Markov Models (HMMs).** By construction, GMMs are static models that do not take into account the dynamics inherent in the speech vectors. In this regard, HMMs [104] are statistical models that capture the temporal dynamics of speech production as an equivalent first-order Markov process. Fig. 4.3 shows an example of a simple HMM which comprises of a sequence of states with a GMM associated with each state. In this example, each state represents a stationary unit of the speech signal also known as "tri-phone". The training procedure for HMMs involves an EM algorithm, where the feature vectors are first temporally aligned to the states using a dy-

namic programming procedure and the aligned feature vectors are used to update the parameters of the state GMM. During the recognition procedure, the most probable sequence of states/phones are estimated (again using a dynamic programming procedure) for a given utterance. The scores generated by each state in the most probable sequence are accumulated to obtain the utterance and speaker specific likelihood. Because the HMMs rely on the phonetic content of the speech signal, they have been dominantly used in speech recognition systems as well as in text-dependent speaker verification systems [123].

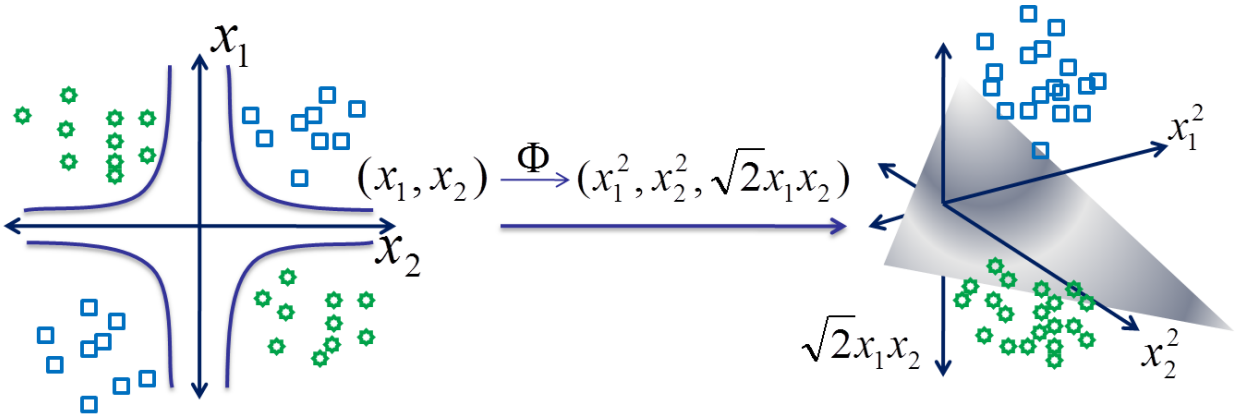
### 4.4.2 Discriminative Models

The discriminative models are optimized to minimize the error on a set of genuine and impostor training samples. They include, among many other approaches, *Support Vector Machines* (SVMs) and *Artificial Neural Networks* (ANNs).

**Support Vector Machines.** SVMs are an attractive choice for implementing discriminative models where they provide good performance in speaker recognition systems even with relatively few data points in the training set and bound on the performance error can be directly estimated from the training data [68]. This is important because of only limited amount of data is usually available for the target speaker. The learning ability of the classifier is controlled by a regularizer in the SVM training, which determines the trade-off between its complexity and its generalization performance. In addition, the SVM training algorithm finds, under general conditions, a unique classifier topology that provides the best out-of-sample performance [68]. The key concept behind an SVM based approach is the use of kernel functions which map the feature vectors to a higher dimensional feature space by using a non-linear transformation $\Phi(.)$. Fig. 4.4(c) illustrates an example of the mapping operation from a two dimensional feature space to a three dimensional




(a)




(b)

Figure 4.4: Discriminative Models: (a) General structure of an SVM with radial basis functions as kernel. (b) Structure of a multi-layer ANN consisting of two hidden layers. (c) An example of a kernel function $K(x, y) = (x.y)^2$, which maps a non-linearly separable classification (left) problem into a linearly separable problem (right) using a non-linear mapping $\Phi(.)$.

space. In the feature space the data points corresponding to the binary classes (denoted by "circles" and "squares") are non-linearly separable. In the higher dimensional space the data points are linearly separable and can be classified correctly by a linear hyper-plane. A binary (two-class) SVM comprises of a linear hyper-plane constructed in the higher dimensional space and is given by

$$f(\mathbf{z}) =< \mathbf{w}, \Phi(\mathbf{z}) > +b$$

where $< ., . >$ defines an inter-product in the higher dimensional space and are the parameters of the hyper-plane. As with SVMs, the hyper-plane parameters $\mathbf{w}$ are obtained as linear expansion over training features $\Phi(\mathbf{x}_n), n = 1, \cdots, N$ as $\mathbf{w} = \sum n = 1N a_n \Phi(\mathbf{x}_n)$ where $a_n$ are the expansion coefficients. Accordingly the inner-products in the expression for $f(\mathbf{z})$ convert into kernel expansions over the training data $\mathbf{x}_n, n = 1, \cdots, N$ by transforming the data to feature space according to

$$\begin{aligned} f(\mathbf{z}) &= < \mathbf{w}, \Phi(\mathbf{z}) > +b & (4.1)\\ &= \mathbf{w} = \sum n = 1N a_n < \Phi(\mathbf{x}_n), \Phi(\mathbf{z}) > +b & (4.2)\\ &= \mathbf{w} = \sum n = 1N a_n \mathbf{K} < \mathbf{x}, \mathbf{z} > +b & (4.3)\\ & & (4.4) \end{aligned}$$

where $K < ., . >$ denotes any symmetric positive-definite kernel that satisfies the Mercer condition and is given by $\mathbf{K} < \mathbf{x}, \mathbf{z} >=< \Phi(\mathbf{x}), \Phi(\mathbf{z}) >$, which is an inner-product in the higher dimensional feature space. For example in Fig. 4.4(c) the kernel function corresponding to $\Phi(.)$ is given by $\mathbf{K}(\mathbf{x}, \mathbf{z}) = (< \mathbf{x}, \mathbf{z} >)^2$. The use of kernel function avoids the curse of dimensionality by avoiding direct inner-product computation in higher-dimensional feature space. Some other

examples of valid kernel functions are radial basis functions $\mathbf{K}(x_i, x_j) = exp(-\sigma(x_i - x_j)^2)$ or polynomial functions $\mathbf{K}(x_i, x_j) = [1 + (x_i.x_j)]^p$. Training of the SVM involves estimating the parameters $a_i$, $b$ that optimizes a quadratic objective function. The exact form of the objective function depends on the topology of the SVM (soft-margin SVM [124], logistic SVM [125] or GiniSVM [166]) and there exist open-source software packages implementing these different algorithms. The following two key steps are the basis for an SVM based recognition:

- *Feature reduction and normalization:* Due to variability in the duration of utterances, the objective of this step is to reduce/equalize the size of the feature vectors to a fixed-length vector. One of the possible approaches could be to use clustering or random selection to determine a pre-determined number of representative vectors. Another approach could use the scores obtained from a generative model (GMM or HMM) as the fixed-dimensional input vector. The features are then scaled and normalized before processed by an SVM.

- *Kernel modeling:* The reduced and normalized feature vectors are used to model each speaker using different types of kernel functions like linear, quadratic, or exponential. For each frame of the feature vector corresponding to the "non-silence" segment of the speech signal, the SVM generates a score and the scores are integrated over the entire utterance to obtain the final decision score. It is important to note that since the scores are required to be integrated it is important that the SVM outputs are properly calibrated. In this regard, logistic SVMs and GiniSVM are useful and have been shown to deliver more robust verification performance compared to traditional soft-margin SVMs. Fi. 4.5 shows an example of SVM based speaker verification system.

**Artificial neural networks (ANNs).** Artificial neural networks [126] have also been used for acoustic recognition systems and are based on discriminant learning. One such example of ANN

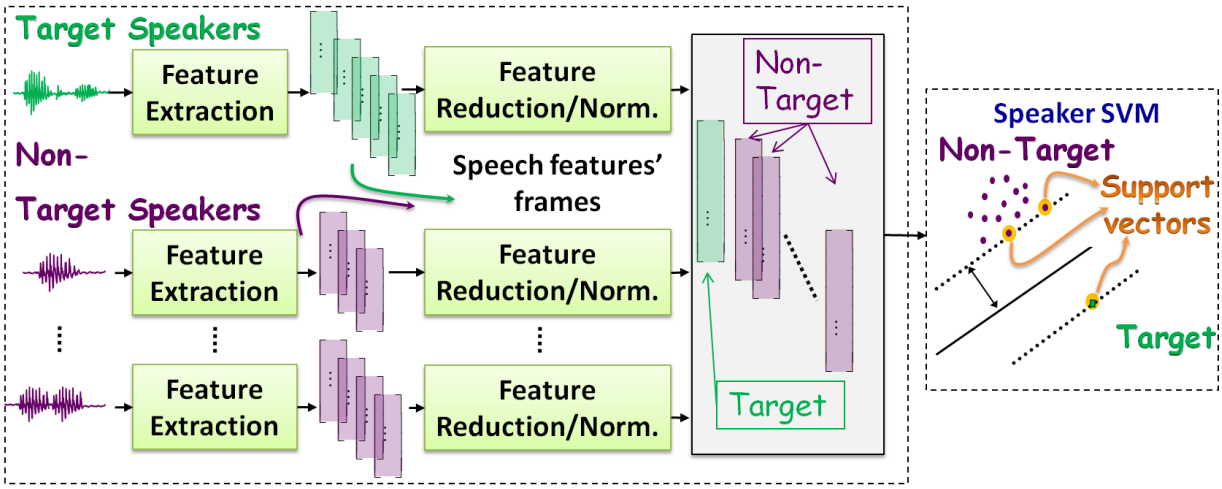



Figure 4.5: Functional architecture of an SVM-based speaker verification system: (left) the extracted features are first aligned, reduced and normalized. The speaker specific and speaker non-specific features are combined to create a dataset used for SVM training. (right) The soft-margin SVM determines the parameter of a hyperplane that separates the target and non-target dataset with the maximum margin.

is the Multilayer Perceptron (MLP) which is a feed-forward neural network comprising of multiple layers and each layer comprising of multiple nodes (as shown in Fig. 4.4(b)). Each node computes a linear weighted sum over its input connections, where the weights of the summation are the adjustable parameters. A non-linear transfer function is applied to the result to compute the output of that node. The weights of the network are estimated by gradient descent based on the back-propagation algorithm. An MLP for speaker verification would classify speaker and impostor's access by scoring each frame of the test utterance. The final utterance score is the mean of the MLP's output over all the frames in the utterance. Despite their discriminate power, the MLP present some disadvantages. The main disadvantage is that their optimal configuration is not easy to select and a lot of data is needed for the training and the cross-validation steps.

**Fusion.** Fusion refers to the process of combining information from multiple sources of evidence to improve the performance of the system. The technique has been also applied in acoustic recognition where a number of different sets of feature are extracted from the speech signal and a

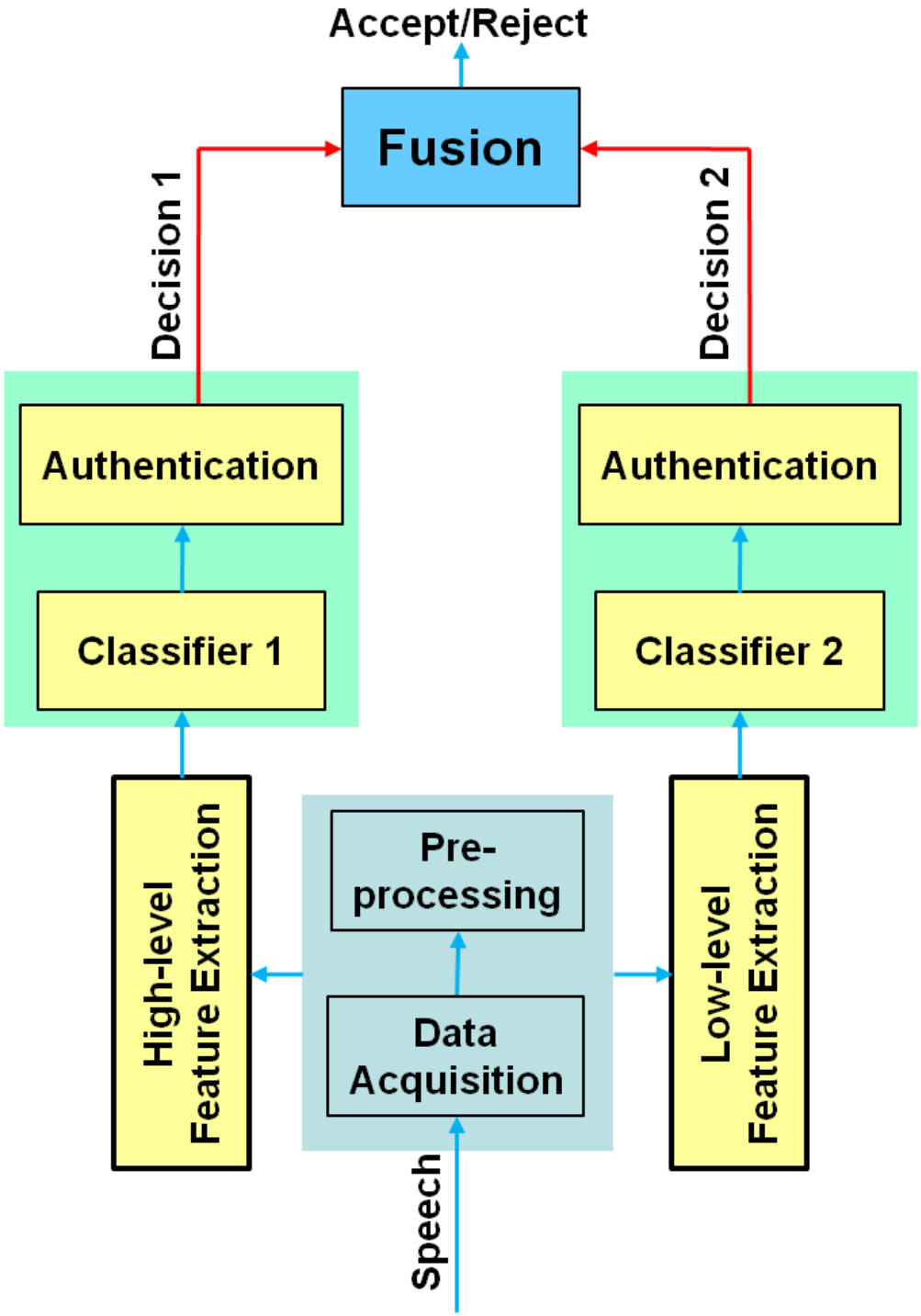



Figure 4.6: An example of fusion of low-level and high-level features for the speaker verification system.

different classifier is trained on each of the feature set. The scores produced by each of the classifier are then combined to arrive at a decision. Ideally, the information contained in the different features should be independent of each other so that each classifier focuses on different regions of the discrimination boundary. Fig. 4.6 shows an example of a fusion technique for speaker verification system that combines "low-level" features like cepstrum or pitch with "high-level" features like prosody or conversational patterns. However, performance gains could also be obtained by fusion of different low-level spectral features (*e.g.,* MFCCs and LPCCs) as they contain some independent spectral information.

**Authentication.** The authentication module in speaker verificaion/recognition systems uses the integrated likelihood scores to determine if the utterance belongs to the target speaker or belongs to an imposter. Mathematically, the task is equivalent to hypothesis testing where given a speech segment $X$ and a claimed identity $S$ the speaker verification system chooses one of the following hypotheses:

$H_S$: $X$ is pronounced by $S$ $H_{\bar{S}}$: $X$ is pronounced by $\bar{S}$ The decision between the two hypotheses is usually based on a likelihood ratio given by

$$\Lambda(X) = p(X|H_S)p(X|H_{\bar{S}}$$

where $p(X|H_S)$ and $p(X|H_{\bar{S}})$ are the integrated likelihood scores (probability density functions) generated by the classifier and $\Theta$ is the threshold to accept or reject $H_S$. Setting the threshold $\Theta$ appropriately for a specific speaker verification application is a challenging task since it depends on environmental conditions like SNR. The threshold is usually chosen during the development phase, and is speaker-independent. However, to be more accurate, the threshold parameter should be chosen to reflect the speaker peculiarities and the inter-speaker variability. Furthermore, if there

is a mismatch between development and test data, the optimal operating point could be different from the pre-determined threshold.

## 4.5 Robust acoustic recognition

The area of acoustic recognition has existed for the last couple of decades but there still exists a large number of challenges that need to be addressed. For example, in the area of speaker recognition/verification the amount of speech data available during enrollment is important in order to have good speaker specific models, especially for generative models like GMMs and HMMs. However, for forensic applications only limited data could be available due to limited access to the target speaker. This was confirmed during the NIST-SRE evaluations [127], where it has been shown that increase in the duration of the utterance improves the recognition performance. Or another challenge in speaker recognition systems is intra-speaker variability. This challenge is as a result of the speaker's voice which could change due to aging, illness, emotions, tiredness and potentially other cosmetic factors and model trained during the training phased might not represent all possible states of the speaker. One of the proposed solutions to this problem is an incremental technique which captures both the short and long-term evolution of a speaker's voice [128].

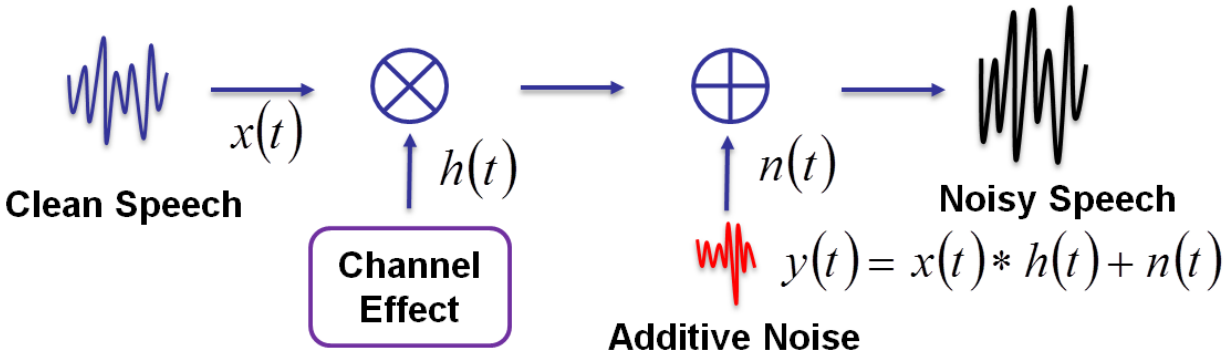
In addition to all the open problems in the area of acoustic recognition, mismatch in training and recognition phases is of more importance as it can limit the application of such systems in real world scenarios. Mismatch in recording conditions during the training and the test/recognition phase pose the main challenge for acoustic recognition systems. Differences in the telephone handset, in the transmission channel and in the recording devices can all introduce variability in recordings and decrease the accuracy of the system. This decrease of accuracy is mainly due to the statistical models that capture not only the speaker characteristics but also the environmental

ones. Hence, the system decision may be biased if the recognition environment is different from the training. A generic framework that models artifacts in a acoustic recognition system is shown in Fig. 4.7, where the sources of interference could either arise due to the additive channel noise or due to the convolutive channel effects. To make speech based recognition systems to be more robust to channel variations, the state-of-the-art systems either use a noise-robust feature extraction algorithm or suitably adapt the models. Fig. 4.7(b) summarizes the approaches that have been used. These approaches in general consist of robust feature extraction techniques and robust modeling. In the following a review of robust feature extraction techniques will be covered as this research proposed technique is in this category.

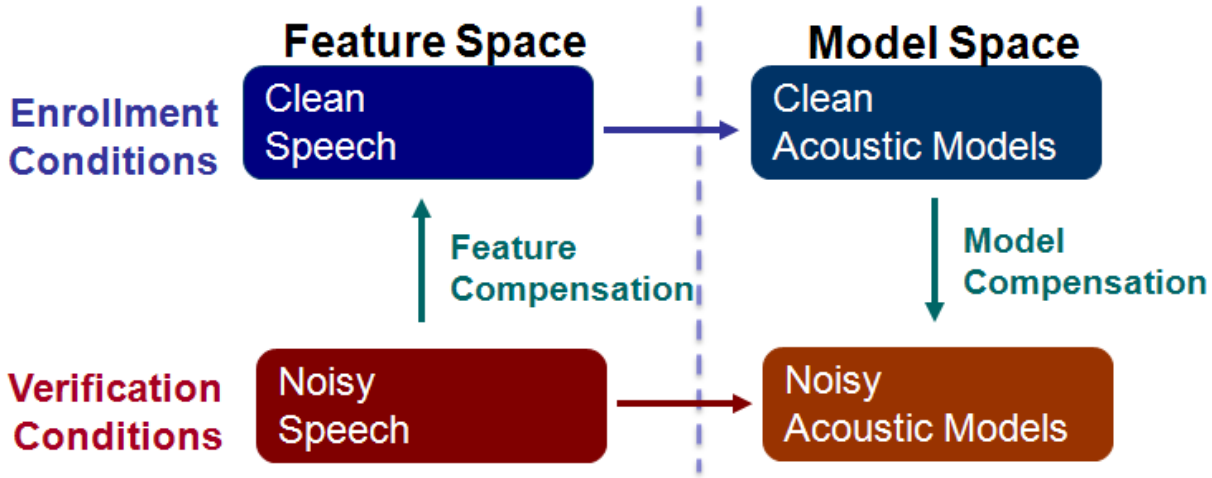
### 4.5.1 Robust Feature Extraction

Different feature based approaches have been proposed to compensate the cross-channel effects which include well-known and widely used techniques such as cepstral mean subtraction (CMS) [59], RASTA filtering [44], and variance normalization [104] as well as recently developed techniques for speaker recognition systems like feature warping [130], stochastic matching [129], and feature mapping [131]. Here will present a brief overview of these techniques:

*Cepstral mean subtraction.* In a Cepstral Mean Subtraction (CMS) method, the mean of cepstral coefficients like MFCC or LPCC computed over a frame of speech is removed from each of the coefficients. The rationale behind CMS is based on the "homomorphic" filtering principles where it can be shown that slow variations in channel conditions are reflected as offsets in the MFCC coefficients. However, CMS is not suitable for additive white noise channel. Also, in addition to mean subtraction sometime the variance of the coefficients is also normalized to improve the noise robustness of the cepstral features.




(a)




(b)

Figure 4.7: (a)Equivalent model of additive and channel noise in a acoustic recognition system (b) Different techniques used for designing robust acoustic recognition systems.

*RASTA filtering.* RASTA (RelAtive SpecTrA) is a generalization of CMS method to compensate the cross-channel mismatch. The method was first introduced to enhance the robustness of speech recognition system and since then it has also been used for speaker recognition systems as well. In RASTA filtering, the low and high frequency components in cepstral coefficients are removed using cepstral band-pass filters.

*Feature warping.* Feature warping was deigned for speaker recognition systems aiming to construct more robust cepstral feature distribution by whitening and hence generating an equivalent normal distribution over each frame of speech. This method delivers a more robust performance than the mean and variance normalization technique, however, the approach is more computationally intensive.

*Feature mapping.* Feature mapping is also designed specifically for speaker recognition systems. The approach is a supervised normalization technique which transforms the channel specific features to a channel independent feature space such that the channel variability is reduced. This is achieved with a set of channel dependent GMMs which are adapted from a channel-independent root model. During the recognition phase, the most likely channel (highest GMM likelihood) is detected, and the relationship between the root model and the channel-dependent model is used for mapping the vectors into channel-independent space.

While the spectral features (MFCC and LPC) accurately extract linear information of speech signals, by construction they do not capture information about nonlinear or higher-order statistical characteristics of the signals, which have been shown to be not insignificant [132, 133]. One of the hypotheses is that many of the non-linear features in speech remain in tact even when the speech signal is corrupted by channel noise. Previous studies in this area have approximated auditory time-series by a low-dimensional non-linear dynamical model. In [133], it was demonstrated

that sustained vowels from different speakers exhibit a nonlinear, non-chaotic behavior that can be embedded in a low dimension manifold of order less than four. Other non-linear speech feature extraction approaches include non-linear transformation/mapping [134, 135], non-linear Maximum Likelihood Feature Transformation [136], kernel based time-series features [137, 138, 139], non-linear discriminant techniques [140], neural predictive coding [141] and other auxiliary methods [142, 143]. we will propose a novel feature extraction technique in which it can extract robust non-linear manifolds embedded in speech signal. The method uses non-linear filtering properties of a functional regression procedure in a reproducing kernel Hilbert space (RKHS). The procedure is semi-parametric and does not make any assumptions on the channel statistics. The hypothesis is that robustness in speech signal is encoded in high-dimensional temporal and spectral manifold which remains intact even in presence of ambient noise. In the following section we will introduce a benchmark setup in order to evaluate our features.

## 4.6 Robust speaker modeling

Several session compensation techniques have been recently developed for both GMM and SVM-based speaker models. Factor analysis (FA) techniques [144] were designed for the GMM-based recognizer and take explicit use of the stochastic properties of the GMM, whereas the methods developed for SVM-based models are often based on linear transformations. One such linear transform based approach uses Maximum Likelihood Linear Regression (MLLR) approach to transform the input parameter of the SVM. MLLR transforms the mean vectors of a speaker-independent model as $\acute{\mu}_k = \mathbf{A}\mu_k + \mathbf{b}$, where $\acute{\mu}_k$ is the adapted mean vector, $\mu_k$ is the world model mean vector and the parameters $\mathbf{A}$ and $\mathbf{b}$ are parameters of the linear transform. $\mathbf{A}$ and $\mathbf{b}$ are estimated by maximizing the likelihood of the training data with a modified EM algorithm. Other normal-

ization techniques for SVMs have also been reported which include nuisance attribute project (NAP) [145, 146] which uses the concept of eigenchannels and withinclass covariance normalization (WCCN) [147, 148] that reweighs each dimension based on different techniques like principal component analysis (PCA). The Nuisance attribute project (NAP) uses an appropriate projection matrix, $\mathbf{P}$ in the feature space to remove subspaces that contain unwanted channel or session variability from the GMM supervectors. The projection matrix filters out the nuisance attributes (e.g. session/channel variability) in the feature space by $\mathbf{P} = \mathbf{I} - \mathbf{U}\mathbf{U}^T$, where $\mathbf{U}$ is the eigenchannel matrix. NAP requires a corpus labeled with speaker and/or session information.

The underlying principle behind factor analysis (FA) when applied to GMMs is the following: When speech samples are recorded from different handsets, the super-vectors or the means of the GMMs could vary and hence require some sort of channel compensation and calibration before they can be compared. For channel compensation to be possible, the channel variability has to be modeled explicitly and the technique that has been used is called joint factor analysis (JFA) [144, 149]. The JFA model considers the variability of a Gaussian supervector as a linear combination of the speaker and channel components. Given a training sample, the speakerdependent and channel (session) dependent supervector $\mathbf{M}$ is decomposed into two statistically independent components as $\mathbf{M} = \mathbf{s} + \mathbf{c}$, where $\mathbf{s}$ and $\mathbf{c}$ are referred to as the speaker and channel (session) supervectors, respectively. The channel variability is explicitly modeled by the channel model of the form $\mathbf{c} = \mathbf{U}\mathbf{x}$ where $\mathbf{U}$ and $\mathbf{x}$ are the channel factors estimated from a given speech utterance and the columns of the matrix $\mathbf{U}$ are the eigen-channels estimated for a given dataset. During enrollment, the channel factors $\mathbf{x}$ are to be estimated jointly with the speaker factors $\mathbf{y}$ of the speaker model of the form $\mathbf{s} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}$ , where $\mathbf{m}$ is the UBM supervector, $\mathbf{V}$ is a rectangular matrix with each of its columns referred to as the eigenvoices and $\mathbf{D}$ is a parameter

matrix of JFA and $\mathbf{z}$ is a latent variable vector for JFA. In this formulation, JFA can be viewed as a two-step generative model which models different speakers under different sessions. The core JFA algorithm comprises the first level and the second or the output level is the GMM generated using the first level. If we consider all the parameters that affect the mean of each component in output GMM, the mean of the session dependent GMM can be expressed as

$$\mathbf{M}_{ki} = \mathbf{m}_k + \mathbf{U}_k\mathbf{x}_i + \mathbf{V}_k\mathbf{y}_{s(i)} + \mathbf{D}_k\mathbf{z}_{ks(i)}$$

with the indices $k$ correspond to different GMM components, $i$ corresponds to session, and $s(i)$ for the speaker in session $i$. The system parameters are $\mathbf{m}_k$, $\mathbf{U}_k$, $\mathbf{V}_k$, and $\mathbf{D}_k$ where $\mathbf{x}_i$, $\mathbf{y}_{s(i)}$, and $\mathbf{z}_{ks(i)}$ are hidden speaker and session variables.

In other approaches the GMM and the SVM principles can be combined to achieve robustness. In [150], the generative GMM-UBM model was used for creating "feature vectors" for the discriminative SVM speaker modeling. For example the mean and the variance of the GMM-UBM states could be used as feature vector for SVM training. When the means of the GMMs are normalized by their variance, the resulting feature vectors are known as supervectors, which have been used in SVM training. The SVM kernel function could be also appropriately chosen that reflects the distance between the pdfs generated by the GMMs. One such measure is the Kullback-Leibler (KL) divergence measure between GMMs. Another extension is the GMM-UBM mean interval (GUMI) kernel which uses a bounded Bhattacharyya distance [151]. The GUMI kernel exploits the speakers information conveyed by the mean of GMM as well as those by the covariance matrices in an effective manner. Another alternative kernel known as probabilistic sequence kernel (PSK) [152] uses output values of the Gaussian functions rather than the Gaussian means to create supervectors. Other SVM approach based on Fisher kernels [125] and probabilistic distance ker-

nels [153] have also been introduced where they use generative sequence models for SVM speaker modeling. Similar hybrid methods have been used for HMMs and SVMs but for applications in speech recognition.

## 4.7 Score normalization

As the name suggests, the score normalization techniques aim to reduce the score variabilities across different channel conditions. The process is equivalent to adapting the speaker-dependent threshold which was briefly discussed in Section 2.3. Most of the normalization techniques used in speaker verification are based on the assumption that the impostors scores follow a Gaussian distribution where the mean and the standard deviation depend on the speaker model and/or test utterance. Different score based normalization techniques have been proposed which includes Znorm [154], Hnorm [155], Tnorm [156], and Dnorm [157]. We describe some of these scores in this section.

*ZNorm.* In zero normalization (ZNorm) technique, a speaker model is first tested against a set of speech signals produced by an imposter, resulting in an imposter similarity score distribution. Speaker-dependent mean and variance normalization parameters are estimated from this distribution. One of the advantages of Znorm is that the estimation of the normalization parameters can be performed offline during the enrollment phase. *TNorm.* TNorm The test normalization (TNorm) is another score normalization technique in which the mean and the standard deviation parameters are estimated using a test utterance. The TNorm is known to improve the performances particularly in the region of low false alarm. However, TNorm has to be performed online while the system is being evaluated. There are several variants of the ZNorm and TNorm that aim to reduce the microphone and transmission channels effects. Among the variants of ZNorm, are the Handset

Normalization (HNorm and the Channel Normalization (CNorm). In the last approach, handset or channel-dependent normalization parameters are estimated by testing each speaker model against a handset or channel-dependent set of imposters. During testing, the type of handset or channel related to the test utterance is first detected and then the corresponding sets of parameters are used for score normalization. The HTNorm, a variant of TNorm, uses basically the same idea as the HNorm. Here, handset-dependent normalization parameters are estimated by testing each test utterance against handset-dependent imposter models. *DNorm.* Both TNorm and ZNorm procedure rely on availability of imposter data. However, when the imposter data is not available an alternate normalization called DNorm can be applied [157] where the pseudo-imposter data are generated from the trained background model using Monte-Carlo techniques.

In the next chapter a novel robust speech feature extraction method will be presented. Both speaker verification and speech recognition results will also be shown in order to present the consistant performance improvement of this new features compared to the conventional methods.

# Chapter 5

# Hierarchical Kernel Auditory Features

This chapter introduces a novel speech feature extraction algorithm using a hierarchical model. This hierarchical model consists of two levels where in the first level the similarity of auditory sensory world is measured with regularized kernel regression technique in a reproducing kernel Hilbert space (RKHS). Then in the second level, the nose-robust features is choosen using a pooling function. The features known as Sparse Auditory Reproducing Kernel (SPARK) are extracted under the hypothesis that the noise-robust information in speech signal is embedded in a subspace spanned by overcomplete and regularized set of gammatone basis functions. Computing the SPARK features involves correlating the speech signal with a pre-computed matrix, thus making the algorithm amenable to DSP based implementation.

## 5.1 Motivation for hierarchical kernel auditory features

Unlike human audition, the performance of speech based recognition systems degrades significantly in the presence of noise and background interference [25, 40]. This can be attributed to inherent mismatch between training and deployment conditions, especially when the characteris-

tics of all possible noise sources are not known in advance. Therefore in literature several strategies have been presented to mitigate this mismatch which can be broadly categorized into three main groups. The first strategy is to improve speech recognition robustness by enhancing the speech signal before feature extraction. Speech enhancement techniques have been designed to improve the perception of speech by objective listeners in noisy conditions or to improve the performance of speech recognition systems. Spectral subtraction (SS) is widely used due to its simpleness which suppress the additive noise in the spectral domain [41]. The second strategy is to make the front-end feature extraction more robust in different conditions. Most of the methods in this group modify the well established technique in order to make them robust. Cepstral mean normalization (CMN) [42] and cepstral variance normalization [43] improve the speech recognition performances by adjusting the feature mean and variance in cepstral domain in order to reduce the convolutive channel distortion. Relative spectra (RASTA) [44] suppress the acoustic noise by high-pass (or band-pass) filtering applied to a log-spectral representation of speech. In recent years, new methods have been proposed to make the exiting features robust by using more advanced signal processing techniques. Examples include feature space nonlinear transformation techniques [45], the ETSI advanced front end (AFE) [46, 162], stereo-based piecewise linear compensation for environments (SPLICE) [47], and power-normalized cepstral coefficients (PNCC) [49]. AFE, for example, integrates several methods to remove both additive and convolutive noises. A two-stage Mel-warped Wiener filtering combined with a SNR-dependent waveform processing is used to reduce the additive noise and a blind equalization is used to mitigate the channel effects. There are some methods in this group which are designed to be inherently robust to mismatched conditions inspiring from human hearing [48]. Recently . The third strategy aims at making the classifier more robust by adjusting or adapting the parameters of the speech models including stochastic

pattern matching methods [50], maximum likelihood estimation (MLE) based signal bias removal method [51], Maximum likelihood linear regression (MLLR) method [52], parallel model combination (PMC) method [53, 54], and joint compensation of additive and convolutive distortions (JAC) based methods [55, 56, 57]. Even though significant improvements in recognition performance can be expected by the application of the third approach, the overall system performance is still limited by the quality of speech features extracted using the second method. Therefore, in this research we focuse on extraction of speech features that are robust to mismatch between training and testing conditions.

Traditionally, speech features used in most of the state-of-the-art speech recognition systems have relied on spectral-based techniques which include Mel-frequency cepstral coefficients (MFCCs) [104], linear predictive coefficients (LPCs) [104, 58, 106], and perceptual linear prediction (PLP) [60]. Noise-robustness is then achieved by modifying these well established techniques to compensate for channel variability. For example, cepstral mean normalization (CMN) [42] and cepstral variance normalization [43] adjust the mean and variance of the speech features in the cepstral domain and in the process reduce the effect of convolutive channel distortion. Another example is the Relative spectra (RASTA) [44] technique which suppresses the acoustic noise by high-pass (or band-pass) filtering of the log-spectral representation of speech. More recently advanced signal processing techniques to improve noise-robustness. These include feature-space non-linear transformation techniques [45], the ETSI advanced front end (AFE) [46, 162], stereo-based piecewise linear compensation for environments (SPLICE) [47] and power-normalized cepstral coefficients (PNCC) [163]. AFE approach, for example, integrates several methods to remove the effects of both additive and convolutive noises. A two-stage Mel-warped Wiener filtering, combined with an SNR-dependent waveform processing is used to reduce the effect of additive noise and a blind

equalization technique is used to mitigate the channel effects. Other attempts in this group has been made to use the auditory models in order to extract features for speech recognition systems. For example, ensemble interval histogram (EIH), an auditory model proposed by [48], has been used as a front-end for speech recognition systems. The EIH is composed of a cochlear filtebank where the output of each filter is attached to a level-crossing detector.

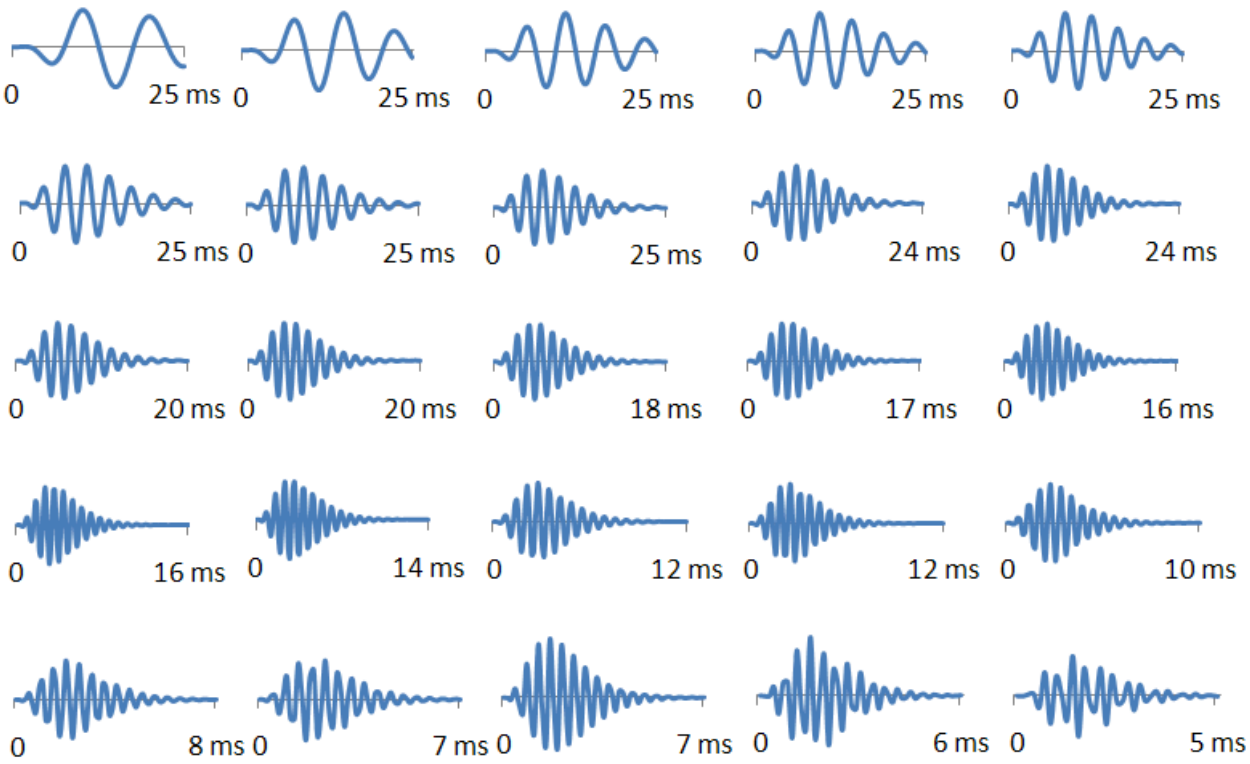
An alternate and more promising approach towards extracting noise-robust speech features is to use data-driven statistical learning techniques that do not make strict assumptions on the spectral properties of the speech signal. Examples include kernel based techniques [166, 167] which operate under the premise that robustness in speech signal is encoded in high-dimensional temporal and spectral manifolds which remain intact even in the presence of ambient noise.

In [167], we had presented a reproducing kernel Hilbert space (RKHS) based regression to extract high-dimensional and noise robust speech features. The procedure required solving a quadratic optimization problem for each frame of speech, thus making the data-driven approach highly computationally intensive. Also, due to its semi-parametric nature, the methods proposed in [166, 167] did not incorporate any a-priori information available from neurobiological or psychoacoustical studies. But, it has been recently demonstrated that cortical neurons use highly efficient and sparse encoding of visual and auditory signals [168, 61, 62]. The study [62] showed that auditory signals can be represented by a group of basis functions which are functionally similar to gammatone functions. Gammatone functions are equivalent to time-domain representations of human cochlear filters and have also been used in psychoacoustical studies [169, 170]. Other studies by a number of auditory neurophysiologists [174, 175, 176] indicates that there is a hierarchical processing in the human auditory cortex where the received signal is first broken down into basic features and later they are integrated into more complex stimuli. These stud-

ies [177, 178] also indicate that the so-called spectro-temporal receptive fields (STRFs) in auditory cortex (AI) can capture different frequencies, spectral scales, and temproral rates. Several researchers have begun to apply these recent developments in neuroscience to speech recognition systems [179, 180, 181, 182, 183]. For example authors in [181] have filtered spectrograms of speech sgnals with spectro-temporal kernels derived from recordings in primary auditory cortex of the ferret. The study [182] presents a model to extract the patch-based features for a word spotting system where a set of patches randomly extracted from the spectrum of training data and in the testing phase a fixed amount of time-frequency flexibility is given to the extracted patches in order to match with the ones from a potential target. The motivation in this research is to apply the kernel based approach [166, 167] to a reproducing kernel Hilbert space (RKHS) spanned by gammatone basis functions and extract sparse, noise-robust, discriminative speech features. The result of incorporating this a-priori information is that SPARK features can be extracted in real-time using pre-computed projection matrices and at the same time demonstrating superior noise-robustness compared to the state-of-the-art features.

## 5.2 Hierarchical architecture

In this section we present two main units of proposed hierarchical architecture in order to generate the speech features. For the analysis presented in this section, we will assume that a window of speech signal is first extracted and the following regression technique is applied on the overlapping windows.




(a)

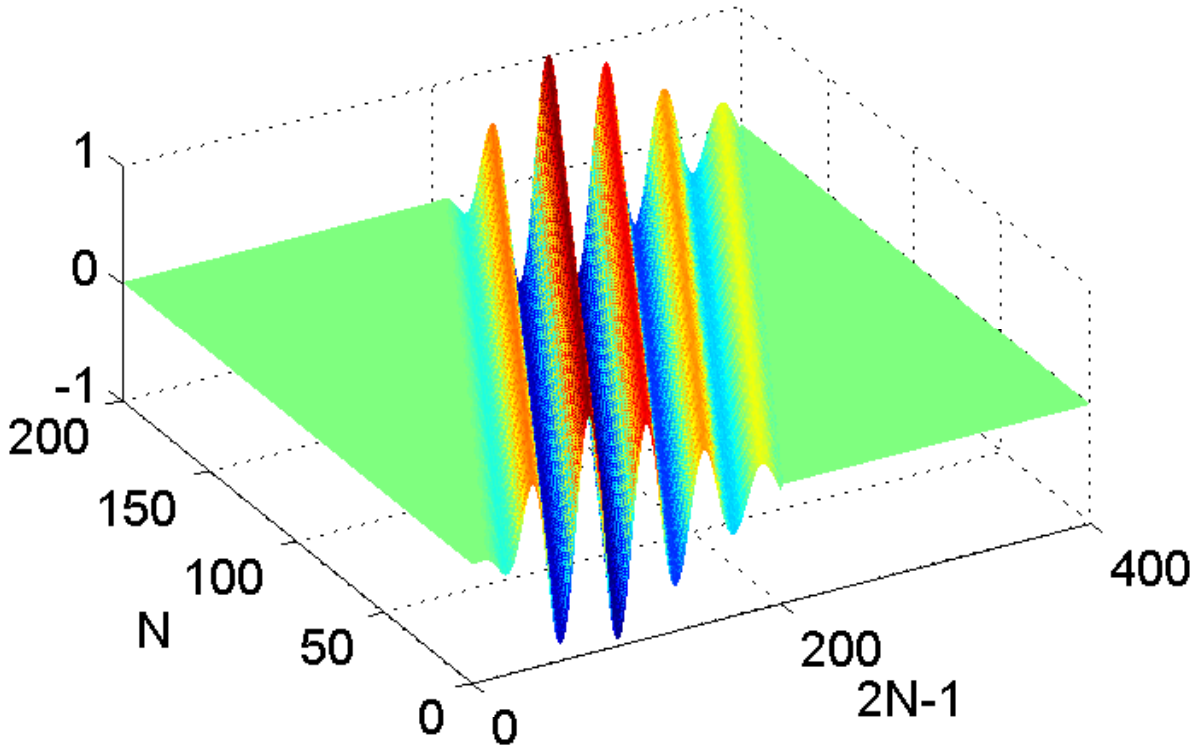
Figure 5.1: A set of 25 gammatone kernel basis functions with center frequencies spanning 100Hz to 4KHz in the ERB space

### 5.2.1 Regularized kernel optimization

Given a stationary, discrete-time speech signal $x[n] \in \mathbb{R}$, where $n = 1, .., N$ denotes the time indices, the objective of the regression is to estimate the parameters of a manifold $f : \mathbb{R}^P \rightarrow \mathbb{R}$ that captures the information embedded in $x[n]$. The function $f$ is assumed to be formed using linear superposition of time-shiftable basis functions $\phi_m$, $m = 1, .., M$ according to

$$f[n] = \sum_{m=1}^{M} \sum_{i=1}^{K} b_{i.m} \phi_m[n - \tau_{i.m}] \tag{5.1}$$

$\tau_{i.m}$ indicates the temporal position of the $i^{\text{th}}$ instance of gammatone function, and $b_{i.m}$ denotes a scaling factor. In this model, all the basis functions and their time-variants are zero padded to




(a)

Figure 5.2: Acyclic convolution matrix $\mathbf{\Phi}_i$ for gammatone basis function $\phi_i$.

have a length of $K$. Note that it has been shown that [61, 62] this shift-invariant mathematical representation can derive the efficient auditory codes through an unsupervised sparse learning. For the purpose of this research we chose the gammatone basis with some fixed parameters as they will be explained later in this section. The reason to choose these basis functions is based on the physiological data in which cochlea exhebits the following characteristics: (a) non-uniform filter bandwidths where each of the frequency resolution is higher at the lower frequency than at the higher frequency, (b) peak gain of the filter centered at $f_C$ decreases as the level of the input increase, and (c) the cochlear filters are spaced more closely at lower frequencies than at higher frequencies. Other reason to choose the gammatone basis functions is that the authors of [61, 62] showed that unsupervised sparse learning over a dataset of natural sounds converges to

the gammatone-shape basis function when the shift-invarient model of (5.1) used to represent the audio signal. Based on the above we chose the gammatone basis functions over the uniform short time Fourier transform (STFT). The gammatone basis functions can be mathematically expressed as cochlear filters given by:

$$\phi_m[n] = a_m n^{k-1} cos(2\pi f_m n) e^{-2\pi b ERB(f_m) n} \tag{5.2}$$

where $f_m$ is the center frequency parameter (Note: gammatone functions are time-domain representation of band-pass cochlear filters), $a_m$ is the amplitude normalization parameter and $k$ is the order of the gammatone basis. The center frequencies are uniformly spaced according to an equivalent rectangular bandwidth (ERB) scale [171]. The parameter $b$ controls the damping ratio of the gammatone basis and is proportional to the ERB of center frequencies. In this research, we have chosen the order $k$ and parameter $b$ to be 4 and 1.019 respectively [172] and the following equation for the ERB$(f_m)$ (suggested by Glasberg and Moore [173]):

$$\text{ERB}(f_m) = 0.108 f_m + 24.7. \tag{5.3}$$

A set of 25 gammatone kernel basis functions with the above parameters are shown in Fig. 5.1. Note that in equation (5.1) a single gammatone basis could occur at multiple times during the speech signal, and hence the model is sufficiently rich to different time-frequency variations.

Equation (5.1) can be expressed in a matrix form as

$$\mathbf{f} = \mathbf{\Phi b} \tag{5.4}$$

and the basis matrix $\Phi$ is defined as

$$\mathbf{\Phi} = [\mathbf{\Phi}_1 \mathbf{\Phi}_2 \cdots \mathbf{\Phi}_M] \tag{5.5}$$

where $\mathbf{\Phi}_i$ is an $N \times K$ acyclic convolution matrix, one for each basis function (see Fig. 5.2). Hence, the matrix $\mathbf{\Phi}$ has dimensions $N \times MK$ and the regression procedure, presented next estimates the parameters $b_{i.m}$. To solve the function regression problem, We first assume that $f$ is an element of a Hilbert space $f \in \mathcal{H}$, where inner-product between two functional elements $f, g \in \mathcal{H}$ will be represented as $\langle f, g \rangle_{\mathcal{H}}$. Note that it can be proved [72] that to every RKHS $\mathcal{H}$ there is a unique positive definite function $K$ called the reproducing kernel of $\mathcal{H}$ that has the reproducing property: $f[n] =< f[l], K[l, n] >$. The function $K$ behaves in $\mathcal{H}$ as the Kronecker's delta function does in $L_2$ [72, 73].

For the purpose of this paper, we take the Hilbert space to be the set of functions of the form defined in equation (5.1) and define the scalar product in this space to be:

$$\langle \sum_{m=1}^{M} \sum_{i=1}^{K} b_{i.m} \phi m[n - \tau_{i.m}], \sum_{m=1}^{M} \sum_{i=1}^{K} c_{i.m} \phi m[n - \tau_{i.m}] \rangle_{\mathcal{H}} \equiv \sum_{m=1}^{M} \sum_{i=1}^{K} b_{i.m} c_{i.m}.$$

Equation (5.6) shows that the norm of the RKHS has the form:

$$||f||^2_{\mathcal{H}_K} = \sum_{m=1}^{M} \sum_{i=1}^{K} b^2_{i.m}, \tag{5.6}$$

The regression involves minimizing the following cost function with respect to $f$

$$\min_{f \in \mathcal{H}} C(f) = \sum_{n=1}^{N} L(x[n], f[n]) \tag{5.7}$$

where $L(.,.)$ is a loss function. While numerous choices of loss functions are possible like the $L_1$ loss function or Vapnik's $\epsilon$-insensitive ($L_\epsilon$) loss function [68], in this paper we have chosen the $L_2$ loss function given by

$$L(x[n], f[n]) = ||x[n] - f[n]||_2^2 \tag{5.8}$$

In the cost function (5.8), we introduce a *stabilizer* or a regularizer $\Omega(f)$ to ensure the solution is more robust. The regularized cost function is given by

$$\min_{f \in \mathcal{H}} H(f) = \sum_{n=1}^{N} L(x[n], f[n]) + \lambda \Omega(f) \tag{5.9}$$

where $\Omega(f)$, in this paper, is chosen as

$$\Omega(f) = ||f||_{\mathcal{H}_K}^2 \tag{5.10}$$

which is a norm in the Hilbert space $\mathcal{H}$ defined by the positive definite function $K$. $||f||_{\mathcal{H}_\mathbf{K}}^2$ determines the smoothness of $f$ based on the regularization parameter $\lambda$ (see the seminal work of [72]). In fact, in [66, 73] it has been shown that when $\mathcal{H}$ is an RKHS defined by specific types of kernels, the use of the regularizer is equivalent to low-pass filtering with cut-off determined by the hyper-parameter $\lambda$.

Cost function (5.9) can be written as

$$J(b_{i.m}) = \sum_{n=1}^{N}(x[n] - \sum_{m=1}^{M}\sum_{i=1}^{K} b_{i.m}\phi_m[n-\tau_{i.m}])^2 + \lambda \sum_{m=1}^{M}\sum_{i=1}^{K} b_{i.m}^2. \quad (5.11)$$

Taking the derivative of (5.11) with respect to parameter $b_{i.m}$ and equating to zero, the following is obtained:

$$\frac{\partial J}{\partial b_{i.m}} = 2\lambda b_{i.m} - \sum_{n=1}^{N} 2\alpha[n]\phi_m[n-\tau_{i.m}] = 0 \quad (5.12)$$

where $\alpha[n]$ is the reconstruction error for speech sample at time instant $n = 1, ..., N$, given by

$$\begin{aligned} \alpha[n] &= x[n] - \sum_{m=1}^{M}\sum_{i=1}^{K} b_{i.m}\phi_m[n-\tau_{i.m}] \\ &= x[n] - f[n] \end{aligned} \quad (5.13)$$

This leads to the minimizer of (5.11) given by

$$b_{i.m} = \frac{1}{\lambda}\sum_{n=1}^{N} \alpha[n]\phi_m[n-\tau_{i.m}]. \quad (5.14)$$

or in a matrix format as

$$\mathbf{b}^* = \frac{1}{\lambda}\mathbf{\Phi}^T\boldsymbol{\alpha}. \quad (5.15)$$

Now $\mathbf{f}$ can be written in terms of $\boldsymbol{\alpha}$ as

$$\mathbf{f} = \frac{1}{\lambda}\mathbf{\Phi}\mathbf{\Phi}^T\boldsymbol{\alpha}. \quad (5.16)$$

Using equations (5.13) and (5.16), (5.11) reduces to

$$\min_{\alpha}(\mathbf{x} - \frac{1}{\lambda}\mathbf{\Phi}\mathbf{\Phi}^T\boldsymbol{\alpha})^T(\mathbf{x} - \frac{1}{\lambda}\mathbf{\Phi}\mathbf{\Phi}^T\boldsymbol{\alpha}) + \frac{1}{\lambda}\boldsymbol{\alpha}^T\mathbf{\Phi}\mathbf{\Phi}^T\boldsymbol{\alpha} \tag{5.17}$$

with the optimum solution written in a matrix format as

$$\boldsymbol{\alpha}^* = \lambda(\mathbf{\Phi}\mathbf{\Phi}^T + \lambda\mathbf{I})^{-1}\mathbf{x}. \tag{5.18}$$

Using equation (5.18), the optimal $\mathbf{b}^*$ also can be calculated as

$$\begin{aligned} \mathbf{b}^* &= \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T + \lambda\mathbf{I})^{-1}\mathbf{x} \\ &= (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^T\mathbf{x}. \end{aligned} \tag{5.19}$$

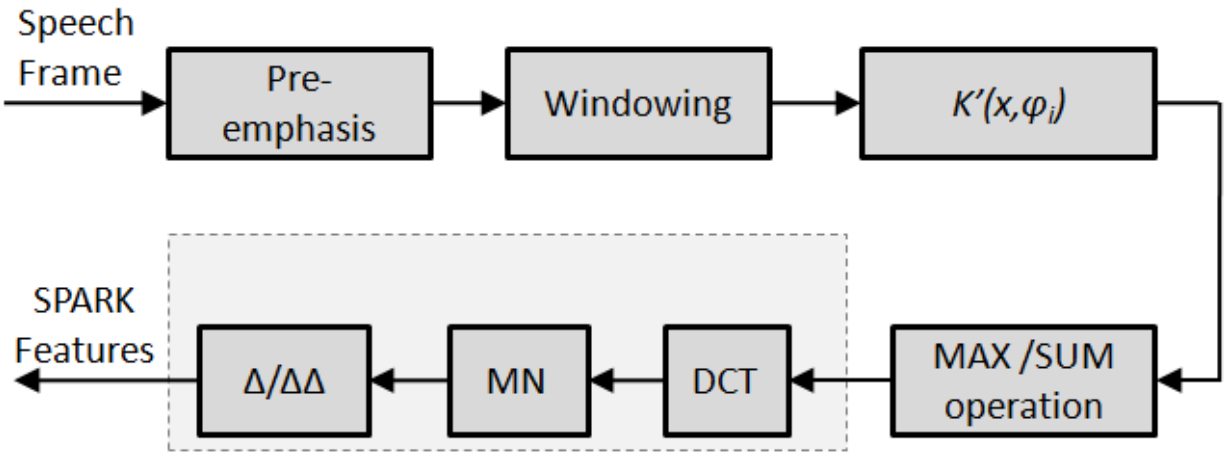
By applying the kernel trick equation (5.19) can be written as

$$\mathbf{b}^* = (K(\mathbf{\Phi}, \mathbf{\Phi}) + \lambda\mathbf{I})^{-1}K(\mathbf{\Phi}, \mathbf{x}) \tag{5.20}$$

The optimal vector $\mathbf{b}^*$ shows the similarity of the input speech signal with each of the basis functions and these parameters will be sent to the second level of this computational model for more complex oparations.

### 5.2.2 Pooling mechanism

An important consequence of projecting the speech signal onto a normalized gammatone function space (representing the STRFs) as shown in equation (5.20) is that the high-energy elements




(a)

Figure 5.3: Signal flow of the SPARK feature extraction

( in $||.||_2$ sense) of the parameter vector $\mathbf{b}$ will capture the salient and the noise-robust aspects of the speech signal in terms of spectral scales, frequencies, and temroral rates. On the other hand, the low-energy components of $\mathbf{b}$ are more susceptible to corruption by noise and they can be eliminated. We therefore apply a weighting function $\zeta(.)$ on elements of $|\mathbf{b}|$ to obtain a more noise-robust representation of the speech signal. This weighting function also emulates the psychoacoustical nonlinear relation between the intensity of sound and its perceived loudness. After nonlinear weighting, a pooling function will choose the winner based on even summation ("SUM") or maximum operation ("MAX"). Note that this pooling mechanism is local and chooses the winner from a set of gammatone function where all of them have a same central frequency. The pooling mechanism we used are:

$$\Psi_m(b_{i.m}) = \zeta\left(\sum_{i=1}^{K} |b_{i.m}|\right), \tag{5.21}$$

$$\Psi_m(b_{i.m}) = \sum_{i=1}^{K} \zeta(|b_{i.m}|), \tag{5.22}$$

and

$$\Psi_m(b_{i.m}) = \max_{i=1}^{K} \zeta(|b_{i.m}|). \tag{5.23}$$

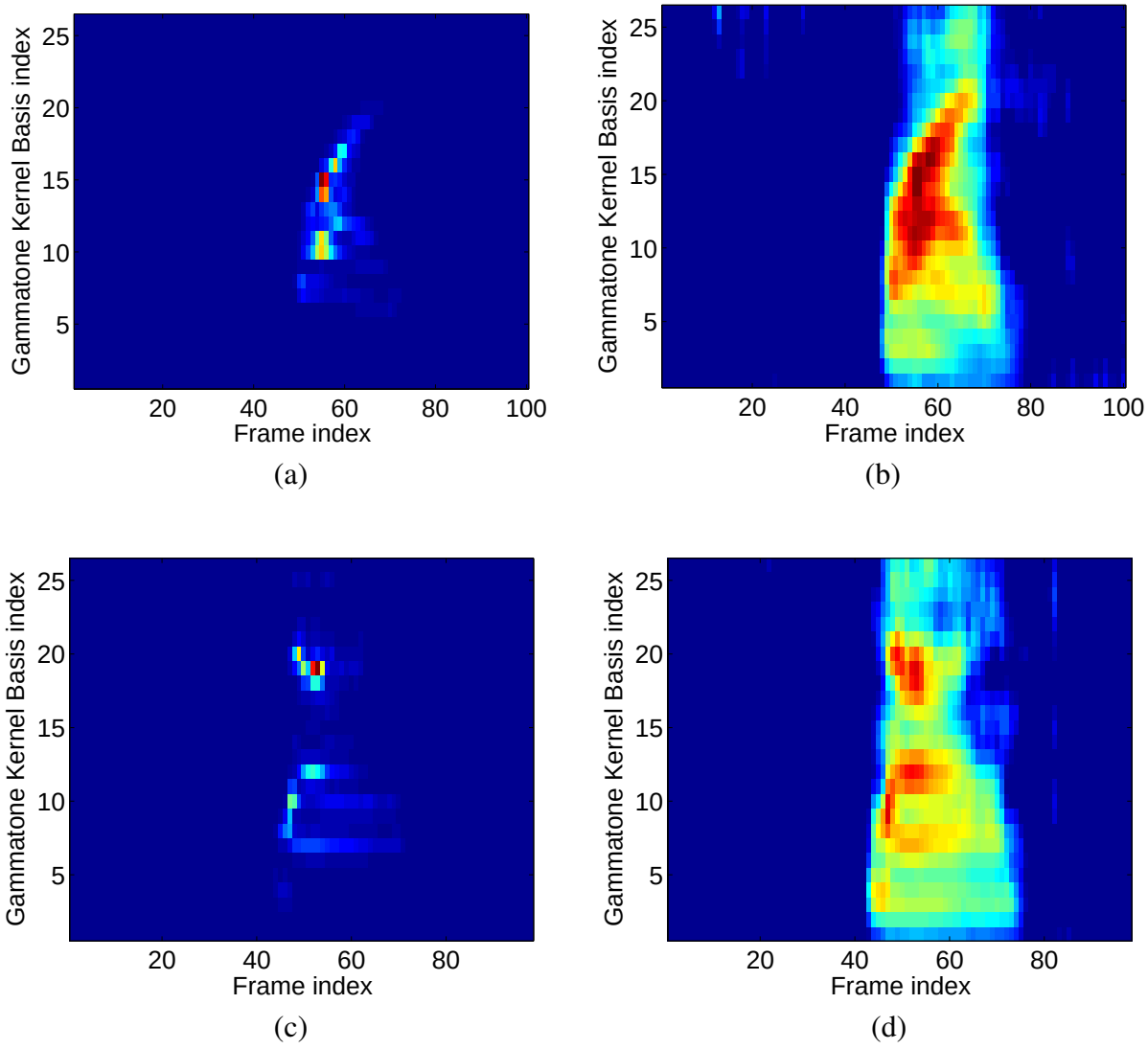



Figure 5.4: Colormaps depicting $\mathbf{b}$ vectors (left column) and IDCT of SPARK feature vectors (right column) obtained for utterances of digit "1" and "9" respectively

## 5.3 Sparse auditory reproducing kernel coefficients

The flow-chart describing the SPARK feature extraction procedure is summarized in Fig. 5.3. The input speech signal is processed by a pre-emphasis filter of the form $x_{pre}(t) = x(t) - 0.97x(t-1)$ after which a 25ms speech segment is extracted using a Hamming window. The parameter vector $\mathbf{b}^*$ is obtained using the kernel regression procedure described in section 5.2.1. A pooling mechanism chooses the parameters that are more robust to noisy conditions. Note that this pooling system also uses a nonlinear weighting function to emulate the psychoacoustical non-linear relation between the intensity of sound and its perceived loudness. Then, a Discrete Cosine Transform (DCT) is applied to de-correlate the features. Like the MFCC based feature extraction, only the first 13 coefficients are used as features. We further apply the mean normalization (MN) to the feature vectors and append the velocity $\Delta$ and acceleration $\Delta\Delta$ parameters to extract a 39 dimension feature vectors for each speech frame. Fig. 5.4 (top row) shows the regression vectors ($\mathbf{b}$) for three different utterances "1" and "9" and bottom row shows the inverse DCT transformation of HKC features. The figures visually depict discriminatory SPARK features for these two different utterances.

## 5.4 Experiments and performance evaluation

In this section two setups is presented for the evaluation of the proposed SPARK features, one for speech recognition system and the other for speaker verification system.

### 5.4.1 Speech recognition setup

In order to compare the speech recognition results with state-of-the-art system reported in literature, we have set up a benchmark system based on the standard Aurora 2 speech recognition task [158].

The setup includes a hidden Markov model (HMM)-based speech recognition architecture, where the speech recognition system is implemented using the hidden Markov toolkit (HTK) package [159]. By the end of the training phase, we have whole word HMM for each digit with 16-state per HMM with three diagonal Gaussian mixture components per state in addition to "sil" and "sp" models.

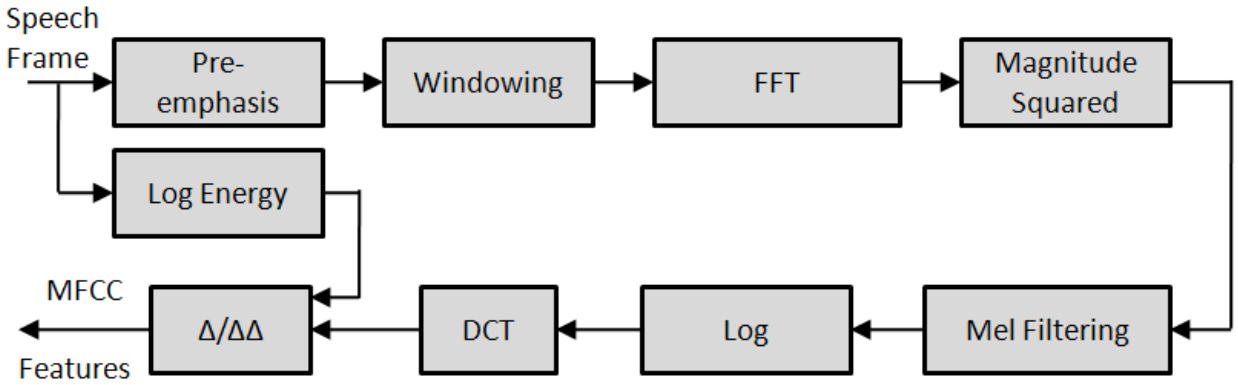
**Aurora 2 database** [158] consists of recognizing English digits in the presence of additive noise and linear convolutional distortion. All the speech data in this database are derived from the TIDigits database at the sampling rate of 8 Khz. The original TIDigits database contains the digit sequences which was originally designed and collected at Texas Instruments Inc. (TI) in 1982. There are 326 speakers in this database with 111 men, 114 women, 50 boys, and 51 girls each pronouncing 77 digit sequences where each speaker group spillited into test and training subsets. The corpus was collected in a quit acoustic environment using an Electro-Voice RE-16 Dynamic Cardiod microphone, digitized at 20 kHz.

In the AURORA 2 database, there are two training mods: training on clean data and multiconditional training on noisy data. The "clean training" corresponds to TIDigits training data downsampled to 8 kHz and filtered with a G712 characteristics. The "multiconditional training" corresponds to TIDigits training data downsampled to 8 kHz and filtered with a G712 characteristics with four different noises added artificially to the data at several SNRs (20 dB, 15 dB, 10 dB, 5 dB, and clean where no noise added), therefore 20 different conditions are taken as input for this

mode.

Three testing sets are provided for the evaluation of the Aurora-2 task. The first testing set (set A) contains 4 subsets of 1001 utterances corrupted by subway, babble, car, and exhibition hall noises, respectively, at different SNR levels (20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB, and clean where no noise added). The second set (set B) contains 4 subsets of 1001 utterances corrupted by restaurant, street, airport, and train station noises at different SNR levels. These distortions have been synthetically introduced to clean (TIDigits) data. The test set C contains 2 subsets of 1001 sentences, corrupted by subway and street noises. The data set C was filtered with the MIRS filter [161] before the addition of noise in order to evaluate the robustness of the speech recognition systems under convolutional distortion mismatch.

The above back-end HMM-based speech recognition system is used with three different feature extraction algorithms for the comparison purpose.




(a)

Figure 5.5: Signal flow of the MFCC feature extraction

**The basic ETSI front-end** [158, 160] is based on Mel Frequency Cepstral Coefficients (MFCCs) which has been widely used in speech based recognition/identification systems [104]. The signal flow of MFCC based feature extraction is shown in Fig. 5.5. The ETSI basic front-end generates the MFCCs with the following parameters. Speech, sampled at 8 kHz, is windowed into frames of

size 200 samples with 80 samples between frames. A logarithmic frame energy measure is calculated for each frame before any processing takes place. Then each frame undergoes pre-emphasis using a filter coefficient equal to 0.97. A Hamming window is then used prior to taking an FFT. Then a magnitude spectrum estimate is used before the filter bank. The basic front-end generates a feature vector consisting of 13 coefficients made up of the frame log-energy measure and cepstral coefficients C1 to C12. In the recognition experiments, velocity and acceleration coefficients are appended to the 13 static features above, to give a total of 39 elements in each feature vector.

Table 5.1: AURORA 2 clean training word accuracy results when ETSI FE is used.

| | Set A | | | | Set B | | | | Set C | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sub | Bab | Car | Exh | Res | Str | Air | Stat | Sub | Str |
| Clean | 99.11 | 98.85 | 98.81 | 99.20 | 99.11 | 98.85 | 98.81 | 99.20 | 99.23 | 98.97 |
| 20dB | 97.11 | 93.65 | 97.55 | 97.04 | 92.45 | 96.37 | 96.00 | 97.35 | 95.58 | 96.13 |
| 15dB | 91.34 | 79.26 | 93.08 | 91.05 | 80.44 | 89.96 | 87.92 | 91.45 | 87.96 | 90.45 |
| 10dB | 74.64 | 54.63 | 74.20 | 73.09 | 59.29 | 67.26 | 66.57 | 70.81 | 70.62 | 72.70 |
| 5dB | 45.47 | 26.96 | 35.55 | 40.64 | 32.05 | 37.03 | 34.92 | 32.09 | 41.76 | 44.32 |
| 0dB | 17.44 | 8.92 | 11.69 | 14.69 | 12.28 | 16.63 | 14.97 | 10.64 | 15.54 | 19.62 |
| -5dB | 8.50 | 2.78 | 8.65 | 8.89 | 4.88 | 8.95 | 8.35 | 7.96 | 8.75 | 9.95 |
| Avg | 61.94 | 52.15 | 59.93 | 60.66 | 54.36 | 59.29 | 58.22 | 58.50 | 59.92 | 61.73 |

Table 5.1 shows the accuracy results of the benchmark speech recognition system on AURORA 2 dataset when ETSI basic front-end (FE) is used. Using this front-end, the avarage word accuracies are $\%58.67$, $\%57.59$, and $\%60.83$ for set a, set b, and set c respectively.

**Conventional gammatone filterbank** uses the auditory gammatone filterbank in order to extract more robust features as shown in Fig. 5.6. In this settings, first a preemphasis of the form $x_{pre}(t) = x(t) - 0.97x(t-1)$ is applied. Then the short-time Fourier transform is performed using Hamming windows of duration 25 ms, with 10 ms between frames, and we used 26 gammatone filters (with the exact gammatone parameters we used in extracting SPARK features). After that, log compression is performed and each speech signal is parameterized with a DCT transformation

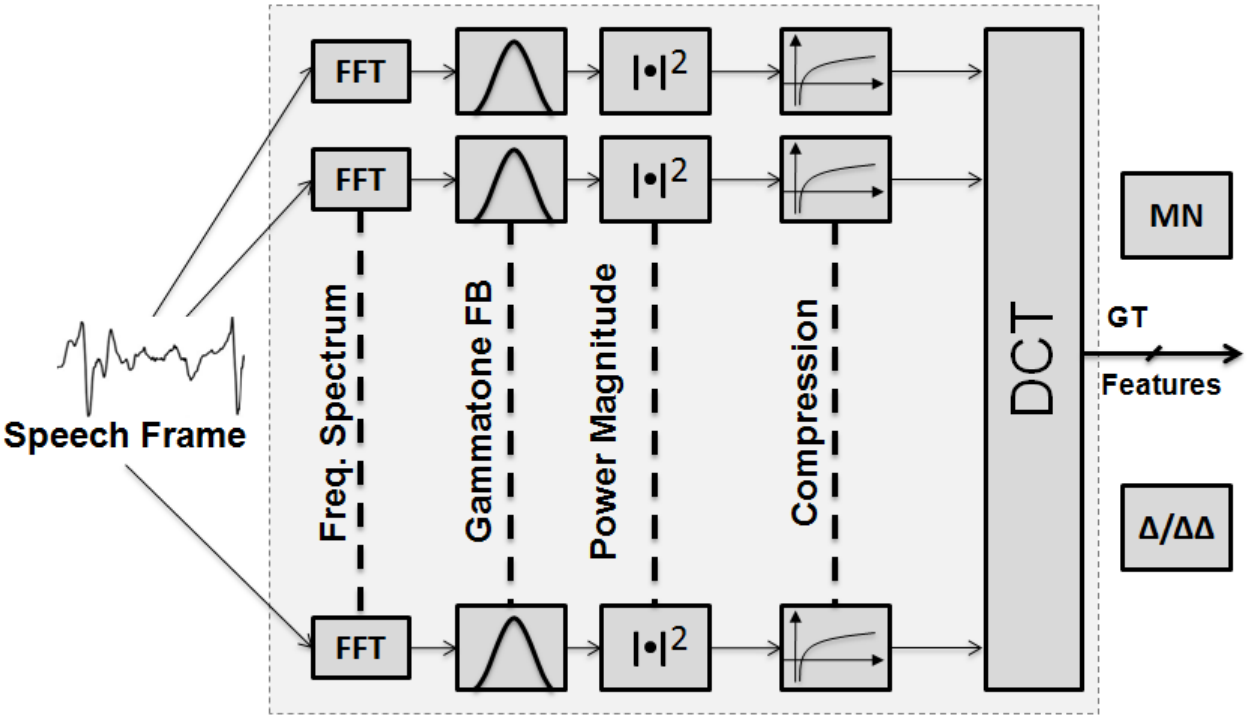



Figure 5.6: Signal flow for the conventional Gammatone filterbank features, note that this figure shows each frame of speech after two steps of pre-emphasis and windowing.

of order 13. The parameters were normalized to have zero mean complemented by their first and second derivatives for a total of 39 coefficients. The results are shown in Fig. 5.6 with the avarage word accuracy of $\%64.53$, $\%67.49$, and $\%65.46$ on set a, set b, and set c respectively.

Table 5.2: AURORA 2 word recognition results when conventional Gammatone filter-bank (GT) is used.

| | Set A | | | | Set B | | | | Set C | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sub | Bab | Car | Exh | Res | Str | Air | Stat | Sub | Str |
| Clean | 99.23 | 99.33 | 98.96 | 99.26 | 99.23 | 99.33 | 98.96 | 99.26 | 99.14 | 99.37 |
| 20dB | 96.62 | 97.94 | 96.99 | 96.67 | 97.97 | 97.58 | 97.67 | 97.81 | 96.90 | 97.49 |
| 15dB | 92.60 | 94.71 | 93.47 | 91.89 | 95.33 | 93.65 | 95.23 | 94.97 | 92.88 | 93.38 |
| 10dB | 79.61 | 83.40 | 78.20 | 77.75 | 85.69 | 82.44 | 86.85 | 83.52 | 80.04 | 81.08 |
| 5dB | 50.26 | 53.08 | 41.57 | 46.37 | 60.12 | 53.02 | 59.23 | 52.92 | 50.60 | 52.09 |
| 0dB | 23.55 | 22.43 | 19.83 | 20.67 | 27.30 | 23.61 | 28.93 | 24.28 | 23.55 | 22.70 |
| -5dB | 14.86 | 12.76 | 12.47 | 12.22 | 12.96 | 12.85 | 15.00 | 13.92 | 14.55 | 12.73 |
| Avg | 65.25 | 66.24 | 63.07 | 63.55 | 68.37 | 66.07 | 68.84 | 66.67 | 65.38 | 65.55 |

**ETSI advanced front-end** is the most recent ETSI standard front-end feature extraction [46, 162]. ETSI advanced front-end (AFE) integrates several methods to remove both additive and convolutive noises. A two-stage Mel-warped Wiener filtering combined with a SNR-dependent waveform processing is used to reduce the additive noise and a blind equalization is used to mitigate the channel effects. The word accuracy results using AFE on the AURORA 2 dataset is shown in Table 5.3.

Table 5.3: AURORA 2 clean training word accuracy results when ETSI AFE is used.

| | Set A | | | | Set B | | | | Set C | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sub | Bab | Car | Exh | Res | Str | Air | Stat | Sub | Str |
| Clean | 99.08 | 99.00 | 99.05 | 99.23 | 99.08 | 99.00 | 99.05 | 99.23 | 99.08 | 99.03 |
| 20dB | 97.91 | 98.31 | 98.48 | 97.90 | 97.97 | 97.64 | 98.39 | 98.36 | 97.36 | 97.70 |
| 15dB | 96.41 | 96.89 | 97.58 | 96.82 | 95.33 | 96.74 | 97.11 | 96.73 | 95.33 | 95.77 |
| 10dB | 92.23 | 92.35 | 95.29 | 92.78 | 90.08 | 92.78 | 93.47 | 93.77 | 90.24 | 90.69 |
| 5dB | 83.82 | 81.08 | 88.49 | 84.05 | 76.27 | 83.28 | 84.07 | 84.57 | 79.03 | 78.17 |
| 0dB | 61.93 | 51.90 | 66.42 | 63.28 | 51.09 | 60.07 | 60.99 | 62.57 | 51.73 | 52.09 |
| -5dB | 30.86 | 19.71 | 30.84 | 32.86 | 18.67 | 29.87 | 28.54 | 29.96 | 24.62 | 25.57 |
| Avg | 80.32 | 77.03 | 82.31 | 80.99 | 75.50 | 79.91 | 80.23 | 80.74 | 76.77 | 77.00 |

**SPARK front-end** We extracted the SPARK features for speech recognition experiments using the procedure described in section 5.3. A 25-ms window with a 10-ms shift has been used and the vector $\mathbf{b}$ has been extracted using 26 kernel gammatone basis functions. In the following experiments the effect of changing different parameters of SPARK features on the performance of the speech recognition system is demonstrated and then a full comparison with the benchmark described above is presented.

In order to reduce the computational complexity of the algorithm, we reduced the size of matrix $\mathbf{\Phi}$ by taking to account different of time-shifts of gammatone basis function.

In order to see the effect of using different kernel functions, we changed the $K$ in equation (5.20). The results presented in Table 5.5 where we used different kernel functions of linear

Table 5.4: The effect of different time-shifts on the SPARK features.

| | Set A | Set B | Set C |
|---|---|---|---|
| SPARK, Shift=3.125 ms | 72.33 | 73.02 | 71.57 |
| SPARK, Shift=4.375 ms | 71.79 | 72.48 | 70.97 |
| SPARK, Shift=7.375 ms | 70.60 | 70.63 | 69.74 |
| SPARK, Shift=11.875 ms | 64.58 | 64.37 | 63.28 |

$K(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y}^T$, exponential $K(\mathbf{x}, \mathbf{y}) = exp(c\mathbf{x}\mathbf{y}^T)$, sigmoid $K(\mathbf{x}, \mathbf{y}) = tanh(a\mathbf{x}\mathbf{y}^T + c)$, and polynomial $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}\mathbf{y}^T)^d$.

Table 5.5: The effect of different kernel functions on the SPARK features.

| | Set A | Set B | Set C |
|---|---|---|---|
| SPARK, Exponential kernel, $c = 0.01$ | 69.83 | 71.45 | 69.52 |
| SPARK, Exponential kernel, $c = 1.0$ | 69.22 | 71.16 | 68.24 |
| SPARK, Sigmoid kernel, $a = 0.01, c = 0$ | 68.35 | 70.60 | 68.89 |
| SPARK, Sigmoid kernel, $a = 0.01, c = -0.01$ | 69.84 | 71.48 | 69.54 |
| SPARK, Linear kernel | 67.80 | 69.65 | 68.30 |
| SPARK, Polynomial kernel, $d = 2$ | 70.77 | 71.14 | 71.07 |
| SPARK, Polynomial kernel, $d = 4$ | 67.89 | 68.24 | 68.05 |

In order to investigate the effect of different pooling mechanisms, we compared the proposed features with different pooling mechanism. First we fixed the pooling function to be $\Psi_m(b_{i.m}) = \zeta\left(\sum_{i=1}^{K} |b_{i.m}|\right)$, and changed the nonlinear weighting function $\zeta(.)$ where the results are presented in Table 5.6 for the polynomial kernel of degree 4 and $\lambda = 0.01$. The results presented in this table clearly show that the non-linear weighting function has a huge effect on the performance of the recognition system.

We ran the same experiments with different pooling function and different kernel function where the results are presented in Tables 5.7 and 5.8. These experiments also show the importance of the non-linear weighting function in extracting SPARK features.

Parameter $\lambda$ controls the smoothness of the regularized regression network presented in sec-

Table 5.6: The effect of different pooling mechanisms (different $\zeta$) when $\Psi = \max\zeta(|\mathbf{b}|)$ and $K(\mathbf{x},\mathbf{y}) = tanh(0.01\mathbf{x}\mathbf{y}^T - 0.01)$.

| | Set A | Set B | Set C |
|---|---|---|---|
| SPARK, $\zeta(.) = (.)^{1/3}$ | 64.91 | 65.60 | 62.60 |
| SPARK, $\zeta(.) = (.)^{1/11}$ | 70.91 | 72.32 | 70.19 |
| SPARK, $\zeta(.) = (.)^{1/13}$ | 70.27 | 71.96 | 69.68 |
| SPARK, $\zeta(.) = (.)^{1/15}$ | 69.83 | 71.24 | 68.88 |
| SPARK, $\zeta(.) = (.)^{1/17}$ | 68.83 | 70.75 | 68.44 |
| SPARK, $\zeta(.) = (.)^{1/19}$ | 68.35 | 70.36 | 68.10 |

Table 5.7: The effect of different pooling mechanisms (different $\zeta$) when $\Psi = \zeta(\sum|\mathbf{b}|)$ and $K(\mathbf{x},\mathbf{y}) = tanh(0.01\mathbf{x}\mathbf{y}^T - 0.01)$.

| | Set A | Set B | Set C |
|---|---|---|---|
| SPARK, $\zeta(.) = (.)^{1/3}$ | 66.39 | 66.50 | 65.59 |
| SPARK, $\zeta(.) = (.)^{1/11}$ | 71.26 | 72.32 | 70.90 |
| SPARK, $\zeta(.) = (.)^{1/13}$ | 70.62 | 72.00 | 70.25 |
| SPARK, $\zeta(.) = (.)^{1/15}$ | 69.84 | 71.48 | 69.54 |

Table 5.8: The effect of different pooling mechanisms (different $\zeta$) when $\Psi = \zeta(\sum|\mathbf{b}|)$ and $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}\mathbf{y}^T)^4$.

| | Set A | Set B | Set C |
|---|---|---|---|
| SPARK, $\zeta(.) = (.)^{1/13}$ | 67.11 | 67.71 | 66.90 |
| SPARK, $\zeta(.) = (.)^{1/15}$ | 67.89 | 68.24 | 68.05 |
| SPARK, $\zeta(.) = (.)^{1/17}$ | 69.06 | 69.01 | 69.34 |
| SPARK, $\zeta(.) = (.)^{1/19}$ | 69.59 | 69.44 | 69.74 |
| SPARK, $\zeta(.) = (.)^{1/25}$ | 70.80 | 70.79 | 70.96 |
| SPARK, $\zeta(.) = (.)^{1/35}$ | 70.90 | 71.87 | 70.50 |

tion 5.2.1. In order to see the effect of this parameter on the speech recohnition performance we ran experiments where we fixed all the parameters except $\lambda$. The results are presented in Table 5.9. As the results show the regularization made the features more robust to the noise in general.

We also compared the SPARK features with the ETSI basic fron-end. Fig. 5.7, 5.8, 5.9, 5.10,

Table 5.9: The effect of $\lambda$ on extracting the SPARK features.

| | Set A | Set B | Set C |
|---|---|---|---|
| SPARK, $\lambda = 0.01$ | 72.33 | 73.02 | 71.57 |
| SPARK, $\lambda = 0.0001$ | 71.41 | 72.35 | 70.25 |
| SPARK, $\lambda = 0.00001$ | 69.18 | 69.73 | 67.99 |
| SPARK, $\lambda = 0.000001$ | 64.12 | 64.79 | 62.68 |

and 5.11 compare the word error-rate obtained by SPARK (with $\lambda = 0.001$) and basic ETSI frond-end based recognizers. The experimental results demonstrate a reduction in the word-error-rate (WER) by $31\%$, $36\%$, and $27\%$ for set A, set B, and set C.

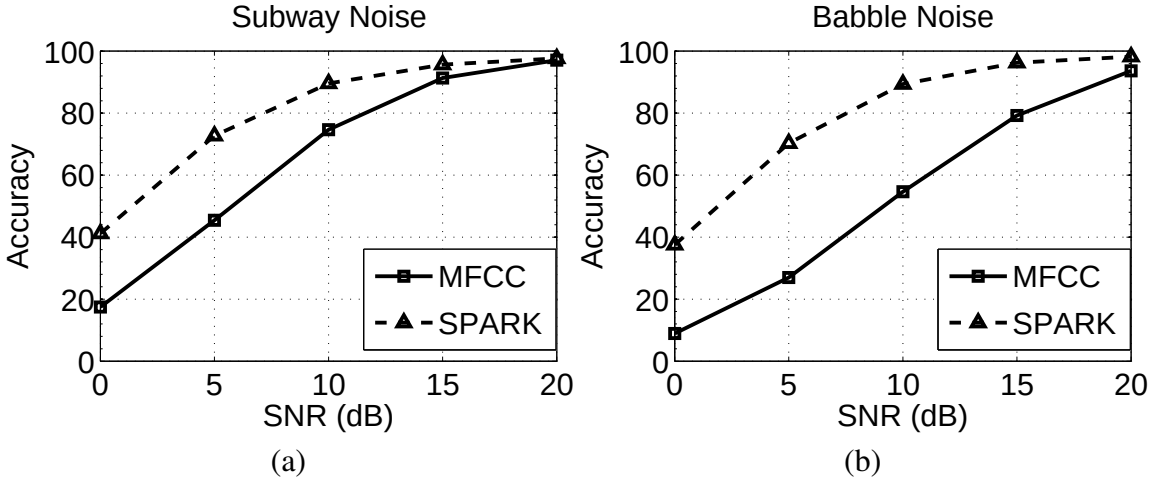
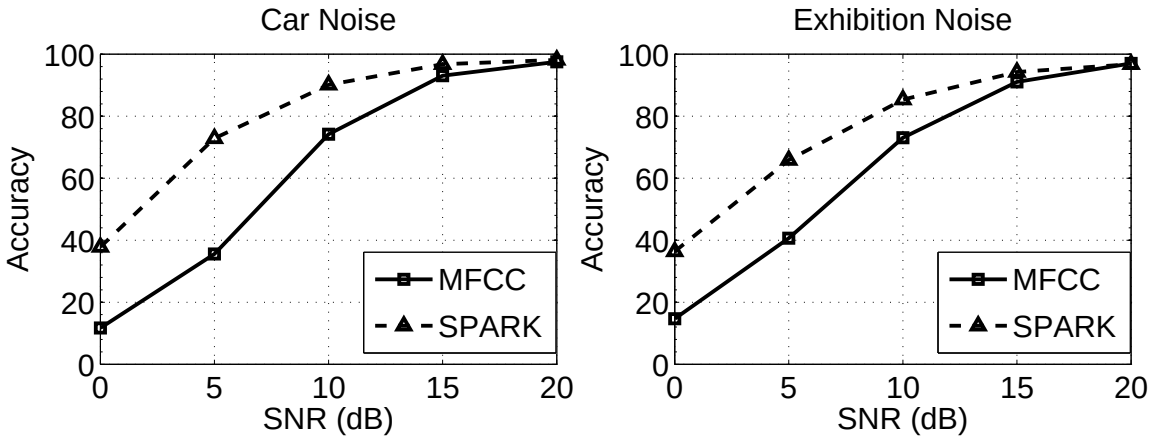


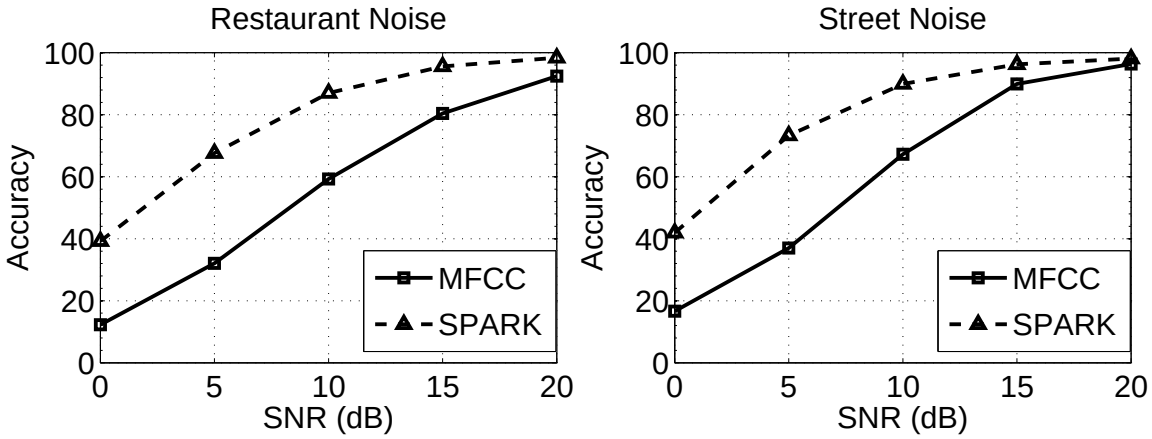

Figure 5.7: Speech recognition accuracy obtained in additive noisy (subway and bable) environments on AURORA 2 database.

We ran another set of experiments to compare the SPARK features to the state-of-the-art ETSI AFE front-end. ETSI AFE uses noise estimation, two-pass Wiener filter-based noise suppression, and blind feature equalization techniques. To incorporate an equivalent noise-compensation to the SPARK features, we used the power bias subtraction (PBS) [163] method. PBS method resembles in some ways to the conventional spectral subtraction (SS), but instead of estimating noise from non-speech parts which usually needs a very accurate voice activity detector (VAD), PBS simply subtracts a bias where the bias is adaptively computed based on the level of the background noise.




(a) (b)

Figure 5.8: Speech recognition accuracy obtained in additive noisy (car and exhibition) environments on AURORA 2 database.

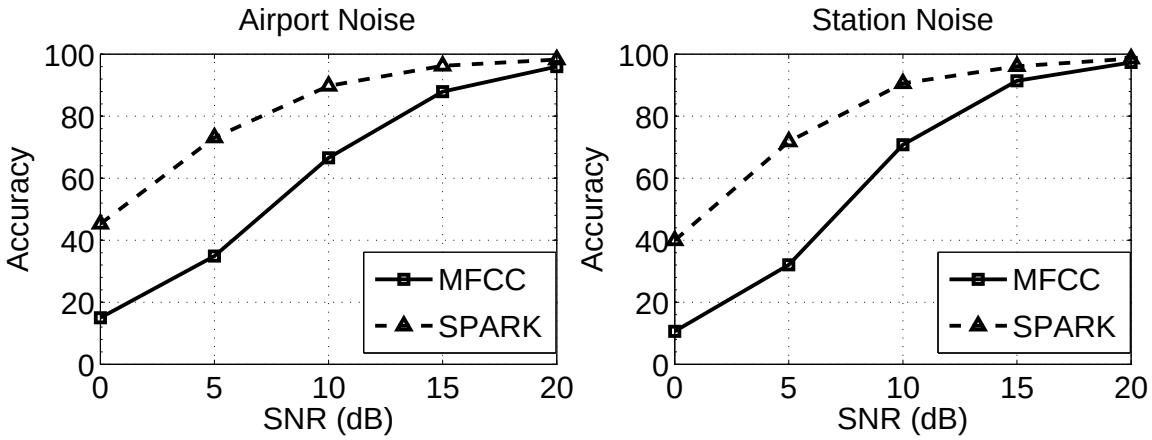



(a) (b)

Figure 5.9: Speech recognition accuracy obtained in additive noisy (restaurant and street) environments on AURORA 2 database.

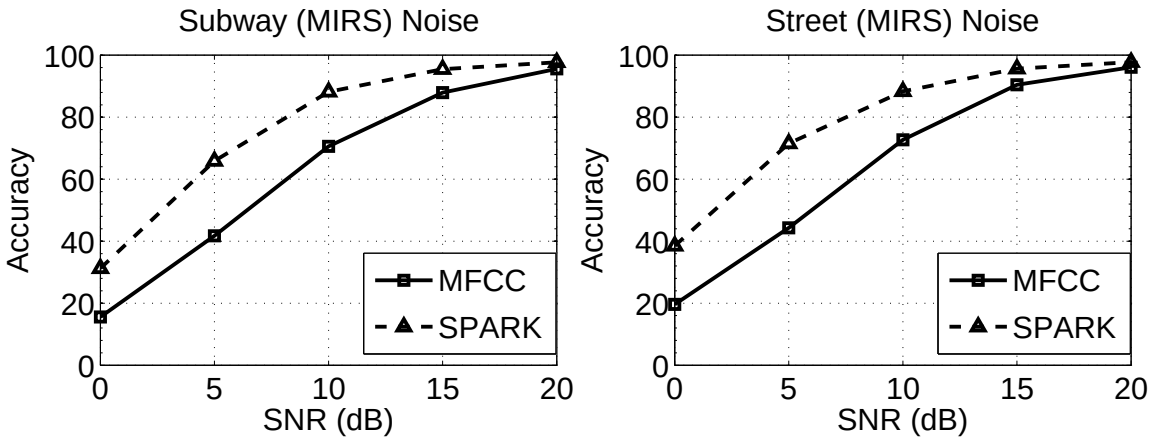
Table 5.10 shows the performance of SPARK+PBS recognition system under different types of noise. These results can be compared to Table 5.3 where they show that SPARK+PBS system consistently performs better than the ETSI AFE all noise types except subway and exhibition noise at low SNR. In fact, SPARK shows an overall relative improvements of 4.69% with respect to the ETSI AFE.

Table 5.11 shows a comparative performance of SPARK+PBS features against other baseline




(a) (b)

Figure 5.10: Speech recognition accuracy obtained in additive noisy (airport and station) environments on AURORA 2 database.




(a) (b)

Figure 5.11: Speech recognition accuracy obtained in different convolutive noisy environments on AURORA 2 database.

systems. The results clearly show that the SPARK+PBS demonstrates improvement over the baseline systems even in clesn condition but the advantage of SPARK+PBS features become more apparent under noisy conditions.

Table 5.10: AURORA 2 word recognition results when SPARK and PBS were used together.

| | Set A | | | | Set B | | | | Set C | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sub | Bab | Car | Exh | Res | Str | Air | Stat | Sub | Str |
| Clean | 99.36 | 99.12 | 99.19 | 99.38 | 99.36 | 99.12 | 99.19 | 99.38 | 99.32 | 99.09 |
| 20dB | 98.10 | 98.70 | 98.69 | 98.15 | 98.83 | 98.37 | 98.90 | 98.58 | 97.82 | 98.04 |
| 15dB | 96.41 | 97.64 | 98.03 | 96.64 | 97.51 | 97.58 | 98.30 | 97.59 | 96.41 | 96.80 |
| 10dB | 92.94 | 95.37 | 95.47 | 92.69 | 94.32 | 94.04 | 96.60 | 95.06 | 92.05 | 93.59 |
| 5dB | 82.87 | 86.61 | 88.76 | 81.67 | 82.99 | 84.22 | 89.41 | 86.76 | 80.60 | 82.98 |
| 0dB | 59.26 | 58.19 | 71.28 | 56.77 | 56.77 | 60.85 | 69.52 | 66.52 | 54.81 | 57.13 |
| -5dB | 27.97 | 21.58 | 34.54 | 25.24 | 21.95 | 27.48 | 32.03 | 33.35 | 25.02 | 25.57 |
| Avg | 79.56 | 79.60 | 83.71 | 78.65 | 78.82 | 80.24 | 83.42 | 82.46 | 78.00 | 79.03 |

Table 5.11: AURORA 2 clean training word accuracy results.

| | Set A | Set B | Set C |
|---|---|---|---|
| ETSI FE WI007 | 58.67 | 57.59 | 60.83 |
| ETSI AFE WI008 | 80.16 | 79.10 | 76.89 |
| Conventional GT | 64.53 | 67.49 | 65.46 |
| SPARK + PBS | 80.38 | 81.24 | 78.52 |

### 5.4.2 Speaker verification setup

For this setup we used a support vector machine (SVM) based speaker verification system in order to discriminate the target speaker from the imposters. This system is based on an open source machine learning software library (Torch [164, 165]). In this system, we used the GMM supervector linear kernel (GSLK) prposed by [150] to measure the dissimilarity between two GMMs, where each GMM obtained by adapting the world model. We used 200 Gussian mixtures for the world model.

**NIST database** Since 1996, the speech group of the National Institute of Standards and Technology (NIST) has been organizing evaluations of text-independent speaker recognition/verification technologies. During the evaluation, a unique data-set and an evaluation protocol are provided to each of the participating research group. The objective is to provide a fair comparison between dif-

ferent speaker verification systems even though the identity of the systems is not publicly revealed.

The effectiveness of the proposed features is evaluated on the NIST 2003 Speaker Recognition Evaluations (SRE) corpus. For the purpose of this work we used the one speaker cellular detection task in NIS SRE 2003 as the evaluation set while for training of the world model (Universal Background Model (UBM)) we used 457 examples from NIST SRE 2000. For evaluation speakers, there were about 2 minutes of speech for training the target speaker models and each test attempt was 15 to 45 seconds long. We only used the male speakers. The evaluation set consists of 149 target speakers. The total number of attempts in the evaluation 17,772 with $10\%$ of true target attempts.

**Evaluation metric** Typically the performance of a speaker verification system is determined by the errors generated by the recognition. There are two types of errors that can occur during a verification task: (a) false acceptance when the system accepts an imposter speaker; and (b) false rejection when the system rejects a valid speaker. Both types of errors are a function of the decision threshold. Choosing a high threshold of acceptance will result in a secure system that will accept only a few trusted speakers, however, at the expense of high false rejection rate (FRR). Similarly choosing a low threshold would make the system more user friendly by reducing false rejection rate but at the expense of high false acceptance rate (FAR). This trade-off is typically depicted using a decision-error trade-off (DET) curve whose example is shown in Fig. 5.12. The FAR and FRR of a verification system defines different operating points on the DET curve. These operating points (shown in Fig. 5.12) vary according to their definition and are considered different performance metrics of the speaker verification system. We describe the commonly used ones below:

Detection Cost Function (DCF): The DCF is a weighted sum of the two error rates and com-

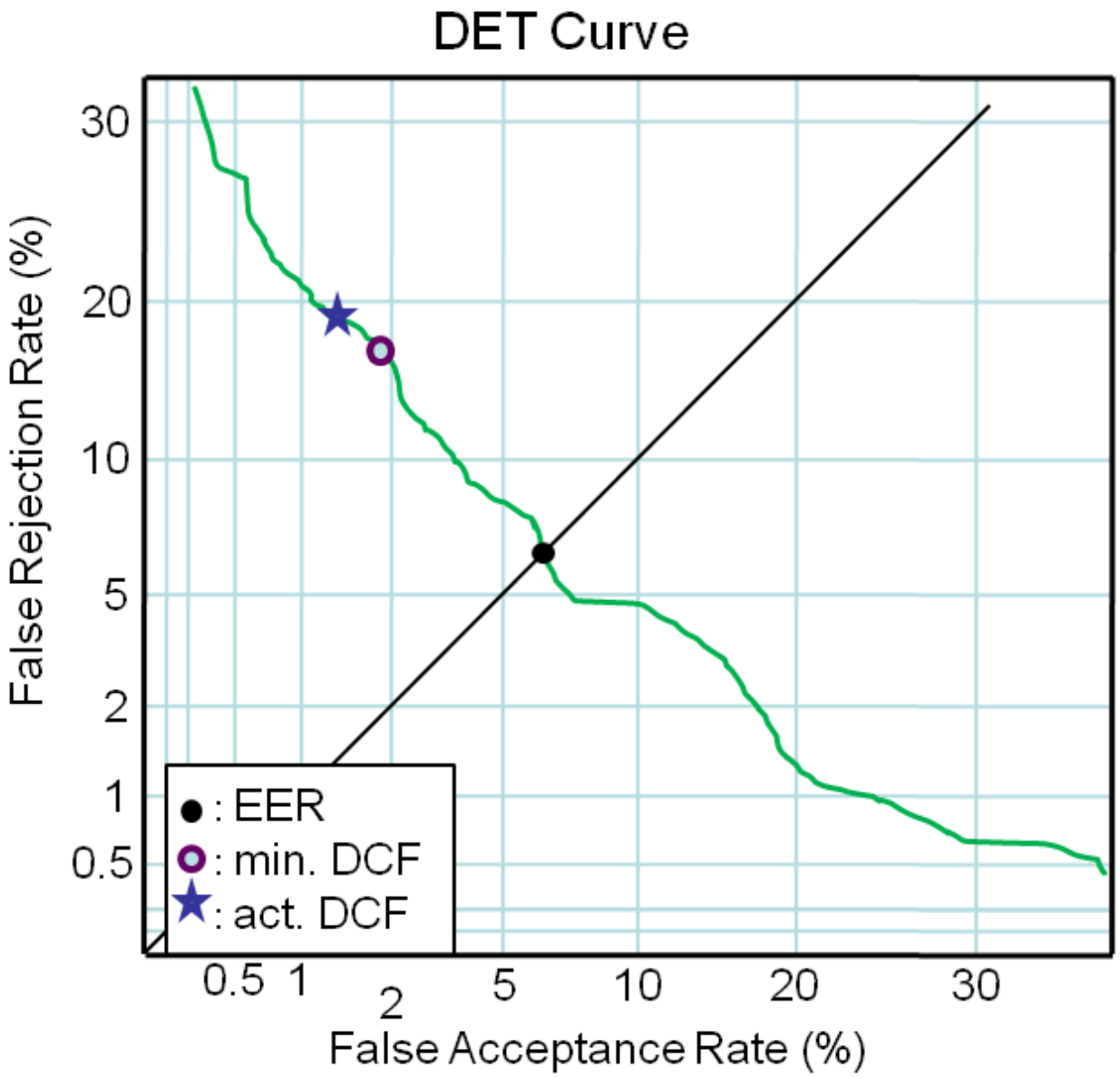



Figure 5.12: An example of DET curve which plots the FRR with respect to FAR.

puted as follows:

$$DCF = (C_{FRR} \times FRR \times P_{Targ}) + (C_{FAR} \times FAR \times (1 - P_{Targ})) \tag{5.24}$$

where $C_{FAR}$ and $C_{FRR}$ denote the cost of false acceptance and cost of false rejection; and

$P_{Targ}$ denotes the prior probability that the utterance belongs to the target speaker. For instance, in evaluations conducted by National Institute of Standards and Technology (NIST), $C_{FAR}$, $C_{FRR}$ ,and $P_{Targ}$ are assumed to be $10$, $1$, $0.01$. Minimum $DCF$ (min. $DCF$) which is the performance metric of the verification system is defined as the smallest value of (5.24) computed over the cross-validation set when the decision threshold is varied. Another related metric is the actual DCF which is the minimum value of (5.24) computed over the test set for the entire range of the decision threshold. An example of the min DCF and actual DCF metric is shown on the DET curve in Fig. 5.12.

Equal Error Rate (EER): An alternative performance measure for speaker verification is the EER which is defined as the FAR which is equal to FRR (see 5.12). Thus, smaller the EER of the system, the superior is the verification system.

As a benchmark system, the verification system described above was developed using MFCC features. MFCCs were extracted for each window of 20ms with 10ms overlap between the adjacent windows. To extract the benchmark features, we used 24 band-pass filters between 300 and 3400HZ. Then each speech signal is parameterized with a DCT transformation of order 16, complemented by the log-energy and their first and second derivatives for a total of 51 coefficients, then all the frames were normalized in order to have a zero mean.

For speaker verification task, we extracted the SPARK features using the procedure described in section 5.3. A 25-ms window with a 10-ms shift has been used and the vector $\mathbf{b}$ has been extracted using 26 kernel gammatone basis functions. Here we kept the first 16 coefficients after the DCT complemented by the first and second derivatives of SPARK features to create a feature vector of 51 coefficients. Fig. 5.13 shows the DET curve comparing the MFCC-CMN features with SPARK features where it clearly demonstrate the effectiveness of the proposed features.

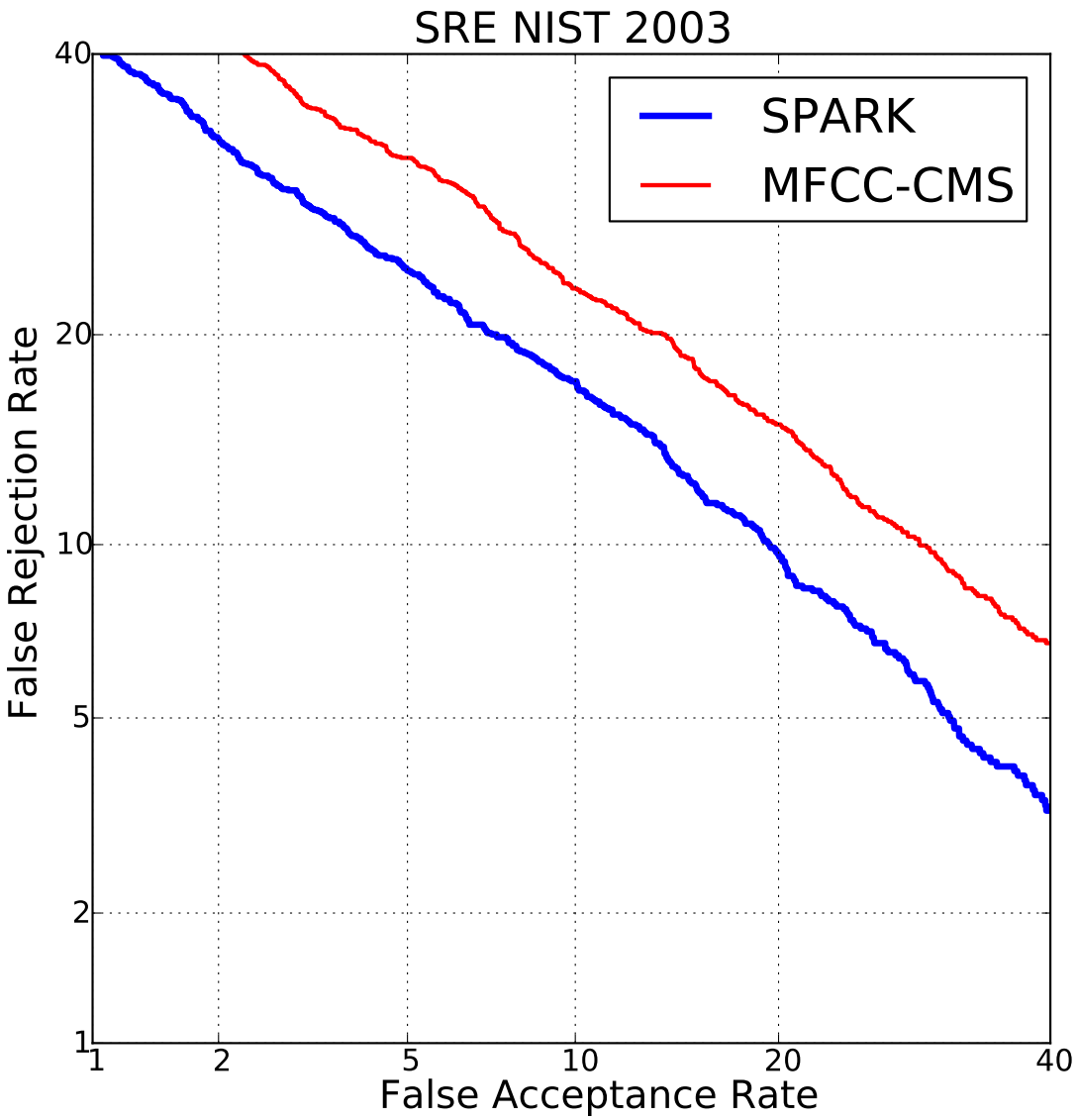



Figure 5.13: DET curve comparing MFCC-CMN and SPARK features.

# Chapter 6

# Concluding Remarks and Future Directions

## 6.1 Summery and concluding remarks

In this work, a miniature acoustic recognition system introduced where the recoding elements are placed in micro/nano scale distance from each other. The mathematical model presented in chapter 2 shows that recording on a miniature microphone array can be approximated with an instantaneous linear mixing model. In chapter 3, a "smart" acquisition system is introduced. At the core of the proposed acquisition system is a min-max optimization of a regularized objective function that yields a sequence of quantized parameters which asymptotically tracks the statistics of the input signals and at the same time removes the cross-correlation of the input space. Therefore, the proposed acquisition system achieves the signal de-correlation along with data conversion at lower digital data bandwidth unlike the conventional data acquisition approach of analog-to-digital conversion followed by data de-correlation process. The performance of this acquisition system is evaluated using synthetic and real recordings and the experiments using the miniature/compact microphone arrays showed a consistent improvement against a standard analog-to-digital converter

for any DSP based source separation algorithms. One of the limitations that prevent a miniature acoustic recognition system to be used in real world applications is its robustness to noisy conditions. It was argued that this issue can be addressed with two general approaches of robust feature extraction techniques and robust modeling. This work also proposed a hierarchical model for robust feature extraction in order to make the miniature acoustic recognition system robust to noisy conditions. The proposed auditory features are extracted in two levels where in the first level of this computational model, the similarity of sensory auditory world is measured through a kernel based approach with a set of gammatone basis functions. The result of incorporating this a-priori information is that these signitures can be extracted in real-time using pre-computed projection matrices. In the second level of this model, the feature are extracted using a pooling mechanism in order to feed into the acoustic recognition unit. The beauty of this approach is its robustness to different noisy conditions and its simplicity in which it can be implemented in real-time using pre-computed matrices therefore it is suitable for the proposed miniature acoustic recognition system.

## 6.2 Future directions

The future work in enhancing the proposed $\Sigma\Delta$ learning for "Smart" acquisition system includes:

- Exploring higher-order noise-shaping $\Sigma\Delta$ modulators for improving the performance of resolution enhancement.

- Extending $\Sigma\Delta$ learning to non-linear signal transforms by embedding kernels into the optimization framework. Incorporating the kernels in signal transformation can capture interesting non-linear information from higher-order statistics of the signal.

- Extending $\Sigma\Delta$ learning to integrate a source separation technique with signal quantization in which the ADC module not only provides the digital representation of the signal but also separates the signal of the interest from other interferences.

In this research framework, $\Sigma\Delta$ learning has been demonstrated to improve the performance of speech based source separation algorithms, but the proposed technique is general and can be applied to any sensor arrays. The potential applications include microphone array hearing aids, microelectrode array in neuroprosthetic devices, miniature radio-frequency antenna arrays and for radar applications.

The future work in extending the proposed hierarchical kernel coefficients includes:

- Learning the basis functions from a speech dataset or updating the gammatone parameters in order to be able to extract more information from the speech signal.

- Exploring other type of basis functions like gammachirp instead of gammatone basis functions.

- Even we introduced a hierarchical just model for the feature extraction module, but this work can be extended to have a hierarchical recognition module as well.

- Exploring the use of hierarchical model for speech and audio coding.

For the speaker verification system, in addition to the features used by the proposed system, there are many other sources of speaker information in the speech signal that can be used. These include idiolect (word usage), prosodic measures and other long-term signal measures. This work will be aided by the increasing use of reliable speech recognition systems for speaker verification research. These high-level features not only offer the potential to improve accuracy, they may also

help improve robustness since they should be less susceptible to channel effects and recent research in this regards show very promising results.

# BIBLIOGRAPHY

# BIBLIOGRAPHY


[1] E. Vilches, I. Escobar, E. Vallejo, and C. Taylor, "Data Mining Applied to Acoustic Bird Species Recognition," in Proc. of International Conference on Pattern Recognition, 2006.

[2] D. J. Mennill, J. M. Burt, K. M. Fristrup, and S. L. Vehrencamp, "Accuracy of an Acoustic Location System for Monitoring the Position of Duetting Songbirds in Tropical Forest," *J. Acoustic Soc. America*, vol. 119, pp. 2832-2839, 2006.

[3] V. M. Trifa, A. N. G. Kirschel, and C. E. Taylor, "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models.," *Journal of Acoustical Society of America*, vol. 123, 2008.

[4] A. N. G. Kirschel, D. A. Earl, Y. Yao, I. A. Escobar, E. Vilches, E. E. Vallejo, and C. E. Taylor, "Using songs to identify individual Mexican antthrush Formicarius moniliger: Comparison of four classification methods," Bioacoustics, vol. 19, 2009.

[5] S. Young and M. Scanlon, "Robotic Vehicle uses Acoustic Sensors for Voice Detection and Diagnostic", Proc. of SPIE, vol. 4024, pp. 72-83, 2000.

[6] S. H. Young and M. V. Scanlon, "Detection and Localization with an Acoustic Array on a Small Robotic Platform in Urban Environments," Progress Report, Army Research Lab , Adelphi, MD, 2003.

[7] C. Clavel, T. Ehrette, and G. Richard, "Events Detection for an Audio-Based Surveillance System," in Proc. of International Conference on Multimedia and Expo, pp. 13061309, 2005.

[8] J. Rouas, J. Louradour, and S. Ambellouis, "Audio Events Detection in Public Transport Vehicle," in Proc. of International Conference on Intelligent Transportation Systems, 2006.

[9] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti "Scream and gunshot detection and localization for audio-surveillance systems," in Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 21-26, 2007.

[10] A. Pikrakis, T. Giannakopoulos and S. Theodoridis, "Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks," in Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008.

[11] V. Cevher, A. C. Sankaranarayanan, J. H. McClellan, R. Chellappa, "Target Tracking using a Joint Acoustic Video System," *IEEE Transactions on Multimedia*, vol. 9, pp. 715-727, June 2007.

[12] H. Zhou, M. Taj, and A. Cavallaro, "Target detection and tracking with heterogeneous sensors,", IEEE Journal of Selected Topics in Signal Processing, vol. 2, issue 4, pp. 503-513, 2008.

[13] K. A. Luthy, "The development of textile based acoustic sensing arrays for sound source acquisition," MS thesis Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, 2003.

[14] E. Grant, K.A. Luthy, J.F. Muth, L.S. Mattos and J.C. Braly, A. Seyam, T. Ghosh, A. Dhawan and K. Natarajan, "Developing portable acoustic arrays on a large-scale e-textile substrate," *International Journal of Clothing Science and Technology*, Vol. 16, pp. 73-83, 2004.

[15] R. Chellappa, G. Qian, and Q. Zheng, "Vehicle Detection and Tracking Using Acoustic and Video Sensors," in Proc. IEEE Intl Conf. Acoustics, Speech, and Signal Processing, pp. 793-796, 2004.

[16] Z. Sun, G. Bebis, and R. Miller, "On-Road Vehicle Detection: A Review," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 694-711, 2006.

[17] A. Morhain and D, Mba, "Bearing defect diagnosis and acoustic emission," in Proceedings of the Institution of Mechanical Engineers, vol. 217, pp. 257-272, 2003.

[18] A. M. Al-Ghamd, and D. Mba, "A comparative experimental study on the use of Acoustic Emission and vibration analysis for bearing defect identification and estimation of defect size," *Mechanical Systems and Signal Processing*, vol. 20, no. 7, 1537-1571, 2006.

[19] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, no. 4861, pp. 1253-1257, 1962.

[20] D. Meuwly, "Voice analysis," in Encyclopaedia of Forensic Sciences, J. A. Siegel, P. J. Saukko, and G. C. Knupfer, Eds., vol. 3, pp. 1413-1421, Academic Press, NY, USA, 2000.

[21] T.G. Clarkson, C.C. Christodoulou, Y. Guan, D. Gorse, D.A. Romano-Critchley, and J.G. Taylor, "Speaker identification for security systems using reinforcement-trained pRAM neural network architectures," IEEE Transactions on Systems, Man and Cybernetics, Part C, vol. 31, no. 1, pp. 65-76, Feb. 2001.

[22] M. Newman, L. Gillick, Y. Ito, D. McAllaster, and B. Peskin, "Speaker verification through large vocabulary continuous speech recognition," in Proc. International Conf. on Spoken Language Processing (ICSLP '96), vol. 4, pp. 2419-2422, Philadelphia, Pa, USA, October 1996.

[23] D. A. Reynolds, R. B. Dunn, and J. J. McLaughlin, "The Lincoln speaker recognition system: NIST EVAL2000," in Proc. International Conf. on Spoken Language Processing (ICSLP '00), vol. 2, pp. 470-473, Beijing, China, October 2000.

[24] L. Wilcox, D. Kimber, and F. Chen, "Audio indexing using speaker identification," in Proc. SPIE Conference on Automatic Systems for the Inspection and Identification of Humans, pp. 149-157, San Diego, Calif, USA, July 1994.

[25] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, pages 261-291, 1995.

[26] D. A. Reynolds, "Automated Speaker Recognition: Current Trends and Future Direction," *Biometrics Colloquium*, vol. 17, June 2005.

[27] G. Davis, Ed., "Noise Reduction in Speech Applications," *CRC Press*, 2002.

[28] S. Kochkin, "MarkeTrak VII: Hearing Loss Population Tops 31 Million People", emphThe Hearing Review, vol. 12, no. 7, pp. 16-29,July 2005.

[29] J. Greenberg and P. Zurek, "Evaluation of an adaptive beamforming method for hearing aids" ,*Journal of the Acoustical Society of America*, vol. 91, pp. 1662-1676, 1992.

[30] B. Widrow, "Microphone Arrays for Hearing Aids," , *IEEE Circuits and Systems Magazine*, vol. 1, no. 2, pp. 2632, 2001.

[31] B. Widrow and F-L Luo, "Microphone Arrays for Hearing Aids: an overview," *Speech Communication*, vol. 39, pp. 139-146, 2003.

[32] M.M. Homayounpour and G. Chollet, "Discrimination of voices of twins and siblings for speaker verification," In Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH), pp. 345348, Madrid, Spain, 1995.

[33] A. Ariyaeeinia, C. Morrison, A. Malegaonkar, and S. Black, "A test of the effectiveness of speaker verification for differentiating between identical twins," Science and Justice, vol. 48, no. 4, pp. 182-186, 2008.

[34] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, New York, 2001.

[35] K-J Koh, J. W. May, G.M. Rebeiz, "A Millimeter-Wave (4045 GHz) 16-Element Phased-Array Transmitter in 0.18-.$\mu$m SiGe BiCMOS Technology, *IEEE Journal on Solid-State Circuits*, vol. 44, no. 5, pp. 1498-1509, 2009.

[36] R.N. Miles and R.R. Hoy, "The development of a biologically-inspired directional microphone for hearing aids," *Audiology and Neuro-Otology*, vol. 11, no. 2, pp. 86-94, 2006.

[37] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, New York, NY: Wiley, 2002.

[38] W. Rudin, *Functional Analysis*, New York: McGraw-Hill, 1991.

[39] J.C. Candy and G.C. Temes, "Oversampled methods for A/D and D/A conversion," *in Oversampled Delta-Sigma Data Converters*, Piscataway, NJ: IEEE Press, pp. 1-29, 1992.

[40] B. H. Juang and T. H. Chen, "The past, present, and future of speech processing," *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 24-48, May 1998.

[41] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pages 113-120, 1979.

[42] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, pp. 13041312, 1974.

[43] S. Molau, F. Hilger, H. Ney, "Feature space normalization in adverse acoustic conditions," In Proc. ICASSP, pp. 656659, 2003.

[44] H. Hermansky and N. Morgan, "Rasta processing of speech," IEEE Trans. Speech Audio Process, vol. 2, no. 4, pages 578-589, 1994.

[45] M. Padmanabhan, S. Dharanipragada, "Maximum likelihood non-linear transformation for environment adaptation in speech recognition systems," In Proc. Eurospeech, pp. 23592362, 2001.

[46] D. Macho, L. Mauuary, B. Noe, Y.M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, F. Saadoun, "Evaluation of a noise robust DSR front-end on Aurora databases," In Proc. ICSLP, pp. 1720, 2002.

[47] L. Deng, A. Acero, M. Plumpe, X. Huang, "Large vocabulary speech recognition under adverse acoustic environments," In Proc. ICSLP, vol. 3, pp. 806809, 2000.

[48] O. Ghitza, "Auditory nerve representation as a basis for speech processing," Edited by S. Furui and M. M. Sondhi, Advances in Speech Signal Process., pp. 453-485, 1992.

[49] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," In Proc. Interspeech 2009, September 2009, Brighton, United Kingdom.

[50] A. Sankar, C-H. Lee, "Robust speech recognition based on stochastic matching," in Proc. ICASSP, vol. 1, pp. 121-124, 1995.

[51] M.G. Rahim, B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Process*, vol. 4, no. 1, pp. 1930, 1996.

[52] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171185, 1995.

[53] M.J.F. Gales, S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," In Proc. ICASSP, pp. 233236, 1992.

[54] M.J.F. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. Thesis, Cambridge University, 1995.

[55] P. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. Thesis, Carnegie Mellon University, 1996.

[56] A. Acero, L. Deng, T. Kristjansson, J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," In Proc. ICSLP, pp. 869872, 2000.

[57] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 975983, 2005.

[58] K. T. Assaleh and R. J. Mammone, "New LP-derived Features for Speaker Identification," IEEE Trans. on Speech Audio Process., vol. 2, no. 4, pages 630-638, 1994.

[59] S. Furui, "Cepstral Analysis Techniques for Automatic Speaker Verification," IEEE Trans. Acoust. Speech Signal Process., vol. 29, pages 254-272, 1981.

[60] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," J. Acoust. Soc. Am., vol. 87, pages 1738-1752, 1990.

[61] E.C. Smith and M.S. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computation*, vol. 17, no. 1, pp. 19-45, 2005.

[62] E.C. Smith and M.S Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978-982, 2006.

[63] M. Riesenhuber, and T. Poggio, "Hierarchical models of object recognition in cortex," Nature Neuroscience, vol. 2, pp. 1019-1025, 1999.

[64] N. Aronszajn, "Theory of reproducing kernels," Trans. American Mathematical Society, vol. 68, pp. 337-404, 1950.

[65] S. Saitoh, "Theory of reproducing kernels and its applications," Longman Scientific and Technical, Harlow, England, 1988.

[66] F. Girosi, M. Jones, and T. Poggio, "Regularization Theory and Neural Networks Architectures," Neural Computation, vol. 7, pages 219-269, 1995.

[67] G. Wahba, "Soft and hard classification by reproducing kernel Hilbert space methods," Proceedings of the National Academy of Sciences, vol. 99, no. 26. pp. 16524-16530, 2002.

[68] V. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995.

[69] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," IEEE Trans. Speech Audio Process., vol. 7, pages 1-17, 1999.

[70] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in Proc. NATO ASI on Speech Production Speech Modeling, vol. 55, pages 241-261, 1990.

[71] B. Scholkopf, C. Burges and A. Smola, eds., Advances in Kernel Methods-Support Vector Learning, MIT Press, Cambridge, 1998.

[72] G. Wahba, "Splines Models for Observational Data," Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990

[73] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization Networks and Support Vector Machines," *Advances in Computational Mathematics*, vol. 13, pp. 1-50, 2000.

[74] H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *Transaction of the A. I. E. E.*, vol. 47, no. 2, pp. 617644, 1928.

[75] C. E. Shannon, "Communication in the presence of noise," *Proceedings of IRE*, vol. 37, pp. 1021, 1949.

[76] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539550, 1999.

[77] L. Bin, T. W. Rondeau, J. H. Reed, and C. W. Bostian, "Analog-to-digital converters," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 6977, 2005.

[78] M.J. Madou, M.J., *Fundamentals of Microfabrication: The Science of Miniaturization*, Boca Raton, FL: CRC Press, 2002.

[79] K.D. Wise, D.J. Anderson, J.F. Hetke, D.R. Kipke, and K. Najafi, "Wireless implantable microsystems: High-density electronic interfaces to the nervous system," *Proceedings of the IEEE*, vol. 92, no. 1, pp. 138-145, Jan. 2004.

[80] C.T. Nordhausen, E.M. Maynard, and R.A. Normann, "Single unit recording capabilities of a 100-microelectrode array," *Brain Research*, vol. 726, pp. 129140, 1996.

[81] K.D. Wise and K. Najafi, "Microfabrication techniques for integrated sensors and microsystems," *Science*, vol. 254, pp. 1335-1342, Nov. 1991.

[82] T. Ajdler and M. Vetterli, "The plenacoustic function, sampling and reconstruction," in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, 2003.

[83] M.N. Do, "Toward sound-based synthesis: the far-field case," in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Canada, 2004.

[84] T. Ajdler, L. Sbaiz, and M. Vetterli, "The Plenacoustic Function and Its Sampling," *IEEE Transaction on Signal Processing*, vol. 54, no. 10, pp. 3790 - 3804, October 2006.

[85] A. Celik, M. Stanacevic, and G. Cauwenberghs, "Gradient flow independent component analysis in micropower VLSI," *Adv. Neural Information Processing Systems (NIPS)*, vol. 8, pp. 187-194, Cambridge: MIT Press, 2006.

[86] J. Barrère, and G. Chabriel, "A Compact Sensor Array for Blind Separation of Sources," *IEEE Transaction on Circuits and Systems: Part I*, vol. 49, no. 5, pp. 565-574, 2002.

[87] E. Oja, "Principal components, minor components, and linear neural networks," *Neural Networks*, vol. 5, no. 6, pp.927-935, 1992.

[88] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, New York, NY: Wiley, 2002.

[89] S. Amari, A. Cichocki, and H.H. Yang, "A new learning algorithm for blind signal separation," *in Adv. Neural Information Processing Systems (NIPS)*, vol. 8, pp. 757-763, Cambridge: MIT Press, 1996.

[90] C. E. Cherry, "Some experiments in the recognition of speech, with one and two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975-979, 1953.

[91] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 18751902, Sep 2005.

[92] P. Földiák and M. Young, "Sparse coding in the primate cortex," In The Handbook of Brain Theory and Neural Networks, pp. 895-898, MIT Press, 1995.

[93] K. P. Körding, P. König, and D. J. Klein, "Learning of sparse auditory receptive fields," In International Joint Conference on Neural Networks, 2002.

[94] J.C. Candy and G.C. Temes, "Oversampled methods for A/D and D/A conversion," *in Oversampled Delta-Sigma Data Converters*, Piscataway, NJ: IEEE Press, pp. 1-29, 1992.

[95] L. Bottou, "Stochastic learning," in Advanced Lectures on Machine Learning, Lecture Notes in Artificial Intelligence, vol. 3176, O. Bousquet and U. von Luxburg, Ed. Berlin: Springer Verlag, 2004, pp. 146-168.

[96] W. Rudin, *Functional Analysis*, New York: McGraw-Hill, 1991.

[97] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434444, 1997.

[98] Z. Koldovský, P. Tichavský and E. Oja, "Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound," IEEE Trans. on Neural Networks, Vol. 17, No. 5, Sept 2006.

[99] E. Vincent, R. Gribonval, and C. Fvotte, "Performance measurement in Blind Audio Source Separation," *IEEE Trans. Audio, Speech, and Lang. Processing*, vol. 14, no. 4, pp. 1462-1469, July 2006.

[100] M. Gupta and S. C. Douglas, "Performance Evaluation of Convolutive Blind Source Separation of Mixtures of Unequal-Level Speech Signals," in Proc. Int. Symposium on Circuits and Systems (ISCAS), New Orleans, Louisiana, 2007.

[101] J. Makhoul, "Linear prediction: A tutorial review," *The Proceedings of the IEEE*, vol. 63, no. 4, pp. 561580, 1975.

[102] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Signal Processing Series, Prentice-Hall Inc., Engelwood Cliffs, New Jersey, 1978.

[103] J.P. Campbell, D. Reynolds, and R. Dunn, "Fusing high- and low-level features for speaker recognition," In Proc. of the European Conference on Speech Communication and Technology (Eurospeech), September 2003.

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

[105] S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4 pp. 357-366, 1980.

[106] S. Furui, "Comparison of speaker recognition methods using static features and dynamic features," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 3, pp. 342-350, 1981.

[107] T. Kinnunen, P. Alku, "On Sepating Glottal Source and Vocal Tract Information in Telephony Speaker Verification," In Proc. of ICASSP, pp. 4545-4548, 2009.

[108] M. Chetouani, M. Faundez-Zanuy, B. Gas, and J.L. Zarader, "Investigation on LP-residual presentations for speaker identification," *Pattern Recognition*, vol. 42, pp. 487-494, 2009.

[109] N. Zheng, T. Lee, and P.C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Sign. Proc. Lett.*, vol. 14, no. 3, pp. 181-184, March 2007.

[110] K.S.R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Sign. Proc. Lett.*, vol. 13, no. 1, pp. 52-55, Jan. 2006.

[111] S.R.M. Prasanna, C.S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Comm.*, vol. 48, pp. 1243-1261, 2006.

[112] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in ICASSP, Las Vegas, pp. 4821-4824, 2008.

[113] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech and Audio Proc.*, vol. 7, no. 5, pp. 569-586, September 1999.

[114] R.E. Slyh, E.G. Hansen, and T.R. Anderson, "Glottal modeling and closed-phase analysis for speaker recognition," in Proc. Speaker Odyssey 2004, Toledo, May 2004, pp. 315-322.

[115] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," in ICASSP 2003.

[116] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, B. Xiang, "Using Prosodic and Conversational Features for High-performance Speaker Recognition: Report from JHU WS'02," in ICASSP 2003.

[117] D. Reynolds, W. Andrews, J. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusávck, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, B. Xiang, "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition," in ICASSP 2003.

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

[119] R. O. Duda , P. E. Hart , D. G. Stork, Pattern recognition, 2nd ed. Wiley-Interscience, New York, 2000.

[120] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol. 39, no. 1, pp. 1-38, 1977.

[121] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech, and Audio Processing, vol. 3, no. 1, pp. 72-83, 1995.

[122] J. Mariéthoz and S. Bengio, "A comparative study of adaptation methods for speaker verification," In Proc. Int. Conf. on Spoken Language Processing (ICSLP), pp. 581-584, Denver, Colorado, USA, 2002.

[123] M. Hébert, "Text-dependent speaker recognition," In Springer handbook of speech processing (Heidelberg, 2008), J. Benesty, M. Sondhi, and Y.Huang, Eds., Springer Verlag, pp. 743-762.

[124] B. Schölkopf and A. Smola., *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge MA, 2001.

[125] T. Jaakkola and D. Haussler, "Probabilistic kernel regression models," Proc. 7th Int. Workshop on Artificial Intelligence and Statistics, 1999.

[126] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan, New York, NY, USA, 1994.

[127] http://www.nist.gov/speech/tests/sre/2008.

[128] C. Fredouille, J. Mariethoz, C. Jaboulet, J. Hennebert, J.-F. Mokbet, and F. Bimbot, "Behavior of a Bayesian adaptation method for incremental enrollment in speaker verification," in Proc. of ICASSP, vol. 2, pp. 1197 - 1200, 2000.

[129] Qi Li, S. Parthasarathy, and Aaron E. Rosenberg, "A fast algorithm for stochastic matching with applications to robust speaker verification," in Proc. ICASSP, pp. 1543-1546, 1997.

[130] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," In Proc. of the IEEE Workshop on Speaker and Language Recognition (Odyssey), June 2001.

[131] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal (ICASSP'03), pp. 53-56, 2003.

[132] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," IEEE Transactions on Speech and Audio Processing, vol. 7, pp. 1-17, 1999.

[133] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," presented at NATO ASI on Speech Production and Speech Modelling, 1990.

[134] H. Hermansky, S. Sharma, P. Jain, "Data-derived nonlinear mapping for feature extraction," Proceedings of the Workshop on Automatic Speech Recognition and Understanding, Dec 1999.

[135] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, H. Hermansky, "Feature Extraction using non-linear transformation for robust speech recognition on the Aurora database", ICASSP 2000

[136] M. K. Omar, M. Hasegawa-Johnson, "Non-Linear Maximum Likelihood Feature Transformation for Speech Recognition" Interspeech, September, 2003.

[137] A. Kocsor and L. Tóth, "Kernel-based feature extraction with a speech technology application," IEEE Trans. On Signal Processing, vol. 52, no. 8, pp.2250-2263, 2004.

[138] H. Huang and J. Zhu, "Kernel based Non-linear Feature Extraction Methods for Speech Recognition," Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06) - Volume 02, pp. 749 - 754, 2006.

[139] A. Lima, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "On the use of kernel PCA for feature extraction in speech recognition," in Proc. Eurospeech, Geneva, Switzerland, pp. 2625-2628, 2003.

[140] Y. Konig, L. Heck, M. Weintraub, K. Sonmez, and R. Esum, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition," Proc. RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications, pp.72-75 (1998).

[141] M. Chetouani, B. Gas, J.L. Zarader, and C. Chavy, "Neural Predictive Coding for Speech Discriminant Feature Extraction : The DFE-NPC," European Symposium on Artificial Neural Networks Bruges (Belgium), 24-26 April 2002.

[142] Q. Zhu, A. Alwan 'Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," Computer Speech and Language, vol. 17, 2003, pp. 381-402.

[143] M. Chetouani, M. Faundez, B. Gas and J.L. Zarader, "Non-linear Speech Feature Extraction for Phoneme Classification and Speaker Recognition," in Nonlinear speech processing : Algorithms and Analysis. Eds. G. Chollet, A. Esposito, M. Faundez, M. Marinaro. Springer Verlag (2005).

[144] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing (ICASSP05), Philadelphia, PA, Mar. 2005, vol. 1, pp. 637640.

[145] A. Solomonoff, C. Quillen, and I. Boardman, "Channel compensation for SVM speaker recognition," in Proc. Odyssey-04 Speaker Language Recognition Workshop, Toledo, Spain, May 2004, pp. 5762.

[146] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process., Philadelphia, PA, Mar. 2005, vol. 1, pp. 629632.

[147] A. O. Hatch and A. Stolcke, "Generalized linear kernels for oneversus-all classification: Application to speaker recognition," In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 585588, France, 2006.

[148] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in Proc. Int. Conf. Spoken Lang. Process., Pittsburgh, PA, Sep. 2006, pp. 14711474

[149] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," technical report CRIM-06/08-14, 2006

[150] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Process. Lett., vol. 13, no. 5, pp. 308311, May 2006.

[151] C. You, K. Lee, and H. Li, "An SVM kernel with GMM supervector based on the Bhattacharyya distance for speaker recognition," IEEE Signal Processing Letters, vol. 16, no. 1, pp. 4952, 2009.

[152] K.-A Lee, C. You, H. Li, and T. Kinnunen, "A GMM-based probabilistic sequence kernel for speaker verification," In Proc. Interspeech, pp. 294-297, Belgium, 2007.

[153] P. Moreno and P. Ho, "A new SVM approach to speaker identification and verification using probabilistic distance kernels," in Proc. 8th Eur. Conf. Speech Commun. Technol., Geneva, Switzerland, pp. 29652968, 2003.

[154] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), vol. 1, pp. 595598, New York, NY, USA, April 1988.

[155] D. A. Reynolds, "The effect of handset variability on speaker recognition performance: experiments on the switchboard corpus," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), vol. 1, pp. 113116, Atlanta, Ga, USA, May 1996.

[156] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," Digital Signal Process., vol. 10, pp. 4254, Jan. 2000.

[157] M. Ben, R. Blouet, and F. Bimbot, "A Monte-Carlomethod for score normalization in automatic speaker verification using Kullback-Leibler distances," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), vol. 1, pp. 689692, Orlando, Fla, USA, May 2002.

[158] H.G Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Proc. ISCA ITRW ASR2000, Paris, France, pp. 181188, 2000.

[159] HTK Speech Recognition Toolkit. Available from: ¡http://htk.eng.cam.ac.uk/¿ (accessed May 2010).

[160] ETSI ES 201 108 Version 1.1.3, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; Front-end feature extraction algorithm; compression algorithms, 2003.

[161] ITU-T Recommendation G.712, "Transmission performance characteristics of pulse code modulation channels," International Telecommunications Union, Geneva, Switzerland, ITU-T Rec.G712, 1996.

[162] ETSI ES 202 050 Version 1.1.5, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; Advanced front-end feature extraction algorithm; compression algorithms," 2007.

[163] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," In Proc. Interspeech, Brighton, United Kingdom, September 2009.

[164] R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: a modular machine learning software library," Technical Report IDIAP-RR 02-46, IDIAP, 2002.

[165] J. Mariéthoz, S. Bengio, and Y. Grandvalet, "Kernel-Based Text-Independent Speaker Verification," in Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods (eds J. Keshet and S. Bengio), John Wiley & Sons, Chichester, UK, 2009.

[166] S. Chakrabartty, Y. Deng and G. Cauwenberghs, "Robust Speech Feature Extraction by Growth Transformation in Reproducing Kernel Hilbert Space," IEEE Transactions on Speech, Language and Acoustics, Vol. 15, no. 6, pp. 1842-1849, 2007.

[167] A. Fazel, S. Chakrabartty, "Non-Linear Filtering in Reproducing Kernel Hilbert Spaces for Noise-Robust Speaker Verification," IEEE International Symposium on Circuits and Systems (ISCAS), Taipei, Taiwan, 2009.

[168] B.A. Olshausen and D.J Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature, vol. 381, pp. 607609, 1996.

[169] R. Patterson, B. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," Frequency Selectivity in Hearing, pp. 123177, 1986.

[170] T. Chi, P. Ru, S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am., vol. 118, no. 2, pp. 887906, 2005.

[171] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," Apple Computer Technical Report, No. 35, 1993.

[172] R. D. Patterson, Holdsworth, I. Nimmo-Smith and P. Rice, "SVOS Final Report: The Auditory Filterbank," APU report, no. 2341, 1988.

[173] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103-108, 1990.

[174] C. M. Wessinger, J. VanMeter, B. Tian, J. Van Lare, J. Pekar and J. P. Rauschecker, "Hierarchical Organization of the Human Auditory Cortex Revealed by Functional Magnetic Resonance Imaging," Journal of Cognitive Neuroscience, vol 13, no. 1, pp. 1-7, 2001.

[175] K. Okada, F. Rong, J. Venezia, W. Matchin, I.-H. Hsieh, K. Saberi, J. T. Serences, and G. Hickok, "Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech," Cereb. Cortex, vol. 20, no. 10, pp. 2486-2495, 2010.

[176] A. Boemio, S. Fromm, A. Braun, and D. Poeppel, "Hierarchical and asymmetric temporal sensitivity in human auditory cortices," Nature Neuroscience, vol. 8, no. 3, pp. 389-395, 2005.

[177] F. Theunissen, K. Sen, and A. J. Doupe"Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds," Journal of Neuroscience, vol. 20, no. 6, pp. 2315-2331, 2000.

[178] J. F. Linden, R.C. Liu, M. Sahani, C. E. Schreiner, and M. M. Merzenich "Spectrotemporal Structure of Receptive Fields in Areas AI and AAF of Mouse Auditory Cortex," Journal of Neurophysiology, vol. 90, no. 4, pp. 2660-2675, 2003.

[179] M. Kleinschmidt, and D. Gelbart, "Improving Word Accuracy with Gabor Feature Extraction," in Proc. ICSLP, 2002.

[180] M. Kleinschmidt, "Localized Spectro-temporal Features for Automatic Speech Recognition," in Proc. Eurospeech, 2003

[181] N. Mesgarani, M. Slaney and S. A. Shamma,“Discrimination of Speech From Nonspeech Based on Multiscale Spectro-Temporal Modulations,” IEEE Transaction on Speech and Audio processing, vol. 14, no. 3, pp. 920-930, 2006.

[182] T. Ezzat, T. Poggio, “Discriminative Word-Spotting Using Ordered Spectro-Temporal Patch Features,” In Proceedings of the 2008 SAPA Workshop, pp. 3540, Brisbane, Australia, September 2008.

[183] J. Bouvrie, T. Ezzat, T. Poggio “Localized Spectro-Temporal Cepstral Analysis of Speech,” in Proc. of ICASSP, Las Vegas, Nevada, 2008.