Speech Enhancement with Applications in Speech Recognition

A First Year Report
Submitted to the School of Computer Engineering
of the Nanyang Technological University

by

Xiao Xiong

for the Confirmation for Admission
to the Degree of Doctor of Philosophy

February 13, 2006

Abstract

The objective of this research is to develop feature compensation techniques to make automatic speech recognition (ASR) systems more robust to noise distortions. The research is important as the performance of ASR systems degrades dramatically in adverse environments, and hence greatly limits the speech recognition application deployment. In this report, we aim to build a generic framework for feature compensation to improve speech recognition accuracy by making speech features less affected by noises.

The degradation of ASR systems under noisy conditions is due to the mismatch between the clean-trained acoustical models and noisy testing speech features presented to the speech recognition engine. Currently, two general approaches are proposed to reduce this mismatch. The first is to adapt the acoustical model to the noisy testing feature, the other is to compensate the noisy testing feature prior to the recognition. We review existing techniques for noise robust speech recognition and find that these techniques generally ignore inter-frame information of the speech signal. We however believe that inter-frame statistics can contribute to noisy speech features compensation and hence propose a vector autoregressive (VAR) model to model speech feature vectors for speech feature reconstruction by either past or future frames prediction. We propose two feature compensation schemes based on the VAR model and the missing feature theory (MFT). Experiments are carried out using the ground-truth data mask on the AURORA-2 database, and our results show significant improvement to recognition accuracy. Specifically, our experimental results showed a relative error rate reductions of 86.51% and 93.9% with respect to the baseline for the subway noise case of test set A and restaurant noise case of test set B at signal to noise ratio equals to -5dB.

The proposed VAR modeling framework is a promising research direction and we will conduct further research to exploit the full potential of this technique.

Acknowledgments

I would like to express my sincere thanks and appreciation to my supervisor, Dr. Chng Eng Siong (NTU), and co-supervisor, Dr. Li Haizhou (I^2R) for their invaluable guidance, support and suggestions. Their knowledge, suggestions, and discussions help me to become a capable researcher. Their encouragement also helps me to overcome the difficulties encountered in my research.

I also want to thank my colleagues in Speech and Dialogue Processing lab of I^2R, for their generous help. I want to thank Ma Bin for his explanation of the HMM, which saved me a lot of time, and Shuanghu, for his generous help on my experiments on speech recognition. I also want to thank George White for helping me adapt the speech recognition engine. My gratitude also goes to Swee Lan, Yeow Kee, Tong Rong, Hendra, Tin Lay, Chen Yu and Boon Pang for their friendship and support.

I am very grateful to the members of our speech team in NTU. It is a pleasure to collaborate with my team mates, Wang Lei, Haishan and Chin Wei.

I am also indebted to my senior graduate fellow Wang Jinjun, for his technical and personal suggestions, especially for his help on HTK training.

Last but not least, I want to thank my family in China, for their constant love and encouragement.

Contents

List of Figures

List of Abbreviation

ASR	Automatic Speech Recognition
BPC	Bayesian Predictive Classification
CASA	Computational Acoustic Scene Analysis
CDCN	Codeword-Dependent Cepstral Normalization
CMN	Cepstral Mean Normalization
CMU	Carnegie Mellon University
CVN	Cepstral Variance Normalization
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EM	Expectation Maximization
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IDCT	Inverse Discrete Cosine Transform
KLT	Karhunen-Loève Transform
LPC	Linear Predictive Coefficients
MAP	Maximum *a posteriori*
MFCC	Mel-Filterbank Cepstral Coefficient
MFT	Missing Feature Theory
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MMSE	Minimum Mean Square Error
PMC	Parallel Model Combination
RATZ	multivaRiate aAussian-based cepsTral normaliZation
SAP	Signal Absence Probability
SNR	Signal to Noise Ratio
STAR	STAtistical Reestimation
STSA	Short-Time Spectral Amplitude
SVD	Singular Value Decomposition
VQ	Vector Quantization
VTLN	Vocal Tract Length Normalization

List of Notation

$*$	Convolution
$\otimes$	Kronecker product
$(\cdot)^T$	Matrix or vector transpose
$\bullet$	Elementary multiplication
$\text{diag}(\mathbf{x})$	Make a diagonal matrix whose diagonal elements are the elements of vector $\mathbf{x}$
$\|\|\mathbf{x}-\mathbf{y}\|\|^2$	Euclidian distance between $\mathbf{x}$ and $\mathbf{y}$
$\|X\|$	Determinant of matrix X
$\mathbf{x} \bullet \mathbf{y}$	Element-wise multiplication of $\mathbf{x}$ and $\mathbf{y}$
$\mathbf{x} \bullet /\mathbf{y}$	Element-wise division of $\mathbf{x}$ by $\mathbf{y}$

Chapter 1

Introduction

The objective of automatic speech recognition (ASR) systems is to recognize the human speeches, such as words and sentences, using algorithms evaluated by a computer without the interference of human. ASR is essentially a statistical pattern recognition task, classifying speech signals into phonemes or words. To create a speech recognition system, a training process that captures the speech statistics using techniques such as hidden Markov model (HMM) [10] is often used.

Currently, state-of-the-art ASR systems can achieve high recognition accuracy under clean acoustic environment [1]. However, under noisy environment, the recognition performance degrades significantly due to the statistical mismatch between the noisy speech feature and the clean-trained acoustic model of the recognition system. The mismatch occurs when the testing condition is different from the training condition, as the acoustic interferences such as additive background noise change the statistics of the speech. It is necessary to address this problem so that the recognition accuracy can be improved to a level which is applicable to real world problems.

The problem of mismatch can be attacked from two approaches: One is the feature compensation approach, i.e. to compensate the noisy speech features prior to the recognition. The other approach, model adaptation approach, adapts the acoustic models of the ASR to the noisy speech feature. Many feature compensation techniques have been proposed, e.g. spectral subtraction [17], Wiener filter [17], feature normalization [48, 78] and model-based estimation of the clean speech features [73],[55]-[57]. These techniques attempt to reduce the effect of the mismatched acoustic environment by estimating the

clean speech features. On the other hand, model adaptation techniques such as parallel model combination (PMC) [58, 75] modifies the distribution of the clean speech to account for the effect of additive noise; maximum likelihood linear regression (MLLR) adaptation [59] techniques transform the means of acoustic model's Gaussians to best fit the noisy observation; maximum *a posteriori* (MAP) [60, 61] adaptation techniques adapt the acoustic model using a Bayesian approach; and STAtistical Reestimation (STAR) [76] technique adds correction terms to mean and variance of the acoustic Gaussians. Because the model adaptation techniques attempt to only match training-testing statistics, their performance can never exceed that of the matched case.

Recently, a new missing feature theory-based (MFT) approach that is inspired by the characteristics of human auditory system attempts to recognize speech using mainly reliable speech features [64]-[72]. The MFT-based techniques usually compensate the corrupted spectral vectors in two steps: the first step is to identify which features of the spectrogram-like time-space representation of the speech[1] are missing, and the second step is to either reconstruct the missing features for recognition [64, 65, 66, 68] or discard them during the recognition process [66, 68]. Because the MFT-based techniques don't make any assumption on noise, they are able to handle various kinds of noise, including non-stationary noise.

One limiting assumption of most of MFT-based techniques is that speech feature vectors of neighbor frames are statistically independent. Although this assumption enables simpler evaluation of the joint probability of the speech feature vectors, they also prohibit the use of the trend information of the speech features in time. For example, in Cooke's [66] state-based imputation method, the missing features are imputed from the acoustical HMM model in the log Mel filterbank domain, i.e., by using the HMM, the independence assumption of the speech features is implicitly applied. In another example, Raj's cluster-based reconstruction of the missing features [65], the log Mel filterbank feature vectors are assumed to be from an independent, identically distributed (IID) multivariate random process and modeled by a Gaussian mixture model (GMM). Raj's method then reconstructs the missing features using the statistics of the trained GMM

[1]For simplicity, we called this representation spectrogram. It is usually in log Mel filterbank domain. The domain of the spectrogram should be clear from the context

with an iterative maximum *a posteriori* (MAP) estimation method. The assumption of IID process disallows the use of inter-frame information.

In another MFT-based technique from Raj, a limited use of inter-frame information is applied in a correlation-based method [65]. In this method, inter-frame statistics are used to reconstruct the missing features by evaluating cross-covariance between two neighboring frames. The correlation method assumes that the speech feature vectors in log Mel filterbank domain are generated from a single wide-sense stationary multivariate process, and the speech feature vectors of every utterance is a realization of the process. This method first captures the cross-covariances statistics of the spectral features during training and then estimates the missing feature using the MAP method during testing. Although inter-frame statistics are utilized, the full potential of the time information in the spectrogram is not exploited. One reason is that the speech signal is very dynamic, and a single wide-sense stationary process is insufficient to model the speech spectrogram.

To fully exploit inter-frame information, we propose to use vector autoregressive model (VAR) to capture the inter-frame statistics for speech feature reconstruction in noisy environment. Although VAR has been used to construct the state distribution of HMM [15], from our survey, it has not been used in the field of feature compensation for noise robust speech recognition. In this report, We use the VAR to capture the relationship between consecutive speech feature vectors. Specifically, the speech feature vector of one frame is represented by a linear combination of the feature vectors of neighbor frames.

To handle the non-stationary characteristics of speech signal, we propose to use multiple VAR models to model the speech feature vectors. The classification of the class is performed by grouping the concatenated speech feature vectors using K-mean algorithm.

Two feature compensation schemes are proposed based on the VAR model and missing feature theory. Experiments has been carried out on the AURORA-2 noisy connected digit database. Results proved the effectiveness our proposed VAR model in exploiting the inter-frame information.

1.1 Report Outline

This report is organized as follows:

Chapter 2 provides a background information on the statistical speech recognition, including the feature extraction, HMM acoustic model and pattern classifier. It also discusses the statistical effect of noise on speech.

Chapter 3 reviews the previous techniques for noise robust speech recognition. For techniques using Bayesian estimation theory to estimate the clean features, a simple derivation of the solution is provided. The connection and difference of the techniques are compared and the relative advantages and weakness of them are analyzed.

Chapter 4 discusses the missing feature theory based techniques. We first introduce the existing MFT-based techniques, with their derivation, followed by the methods for generating data masks.

In Chapter 5, we propose the VAR for modeling the speech feature vectors. Two feature compensation schemes is proposed in the MFT framework and the experimental results are discussed.

Finally, we conclude in Chapter 6, where we also discuss about the directions and schedule of our future research.

Chapter 2

Statistical Speech Recognition

In this chapter, we first review the fundamental aspects of speech recognition, the modules and operation of a speech recognition system. Next, we discuss the effects of noise on the performance of speech recognition systems.

2.1 Modules of Speech Recognition System

Speech recognition is a pattern classification problem: "The goal is to take one pattern, the speech signal, and classify it as a sequence of previously learned patterns, e.g., words or subword units such as phonemes."[1] A state-of-the-art speech recognition system consists of six modules, i.e. the feature extraction, acoustic model, language model, word lexicon, pattern classifier and the confidence scoring module (See Fig.2.1). The feature extraction module extracts useful speech information from the raw speech samples for classification. The acoustic model, the language model and the word lexicon are used to capture the speech information in different levels. Specifically, the acoustic model captures the information in the feature level to facilitate the pattern classification, the word lexicon specifies all the correct words and the language model ensures that the recognized sentence is syntactically correct. The pattern classifier, the heart of a speech recognition system, makes use of all the available information to make the best guess about the underlying sentence of the input speech signal, and finally the confidence scoring module is used to verify the recognized sentence.

In the following sections, we discuss these six modules and pay particular attention to the feature extraction module due to its close relationship with the feature compensation techniques.

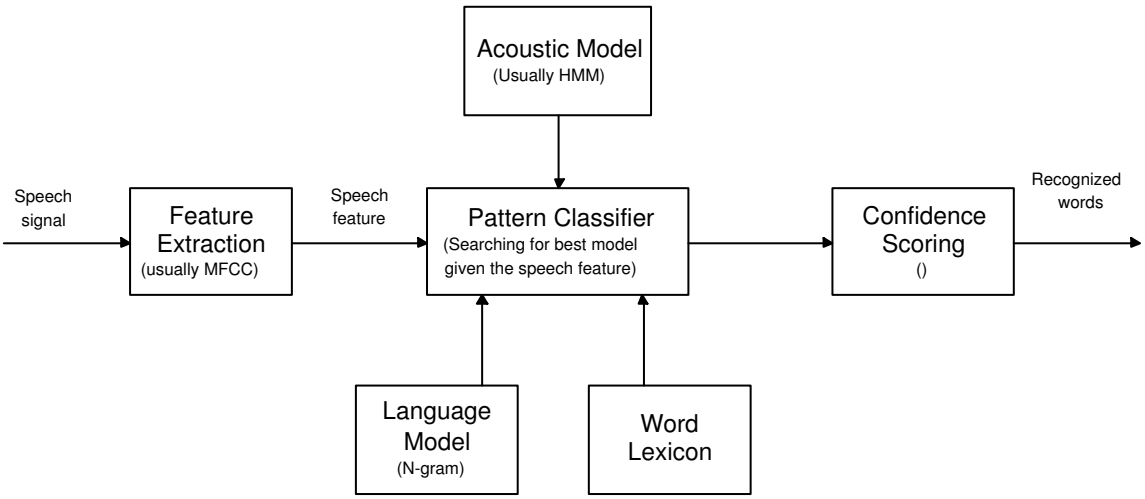

Figure 2.1: The six modules of a modern ASR system (adapted from [1]).

2.1.1 Feature Extraction

The first step of a speech recognition process is to extract the speech features from the speech signal. The purpose of feature extraction is two fold: first is to compress the speech signal to features; second is to use features that are insensitive to speech variation and changes of environmental conditions. The second reason is important as speech recognition accuracy degrades significantly when speech variations occur and environment condition changes. Some examples of these variations include accent differences, male-female vocal tract difference etc. Examples of environment condition changes include changes in the transmission channel, changes in characteristics of the microphone, the background noise, and cocktail effects, etc. To make the speech recognition robust, the speech feature should be designed to be insensitive to these variations and changes. For speaker-independent ASR systems, the feature is generally designed to be sensitive to the difference between words or phonemes, and be insensitive to the differences between persons and between various acoustic environments. Currently, the most popular speech feature is the Mel filterbank cepstral coefficients (MFCC). MFCC feature vector is usually a 39 dimensional vector, consisting of 13 basic features, and their first and second derivatives. A good overview of feature extraction techniques is available in [35]. The procedure of feature extraction is summarized as follows (see Fig.2.3):

DC offset removal and pre-emphasis This module removes the DC offset of the speech signal and pre-emphasis the signal spectrum by approximately 20 dB per

decade to flatten the spectrum of the speech signal. The pre-emphasis filter is used to offset the negative spectral slope of voiced speech signal to improve the efficiency of the spectral analysis [35].

Framing Human speech signal is slowly time varying and can be treated as a stationary process when considered under a short time frame. Therefore, the speech signal is usually separated into small duration blocks, called frames, and the spectral analysis is performed on these frames. The neighboring blocks are overlapped by 1/2 to 2/3 length of the frame and the frame shift is the frame length minus the frame overlap. The commonly used frame length and frame shift are 20-30 ms and 10 ms respectively for speech recognition task because the positions of the articulators do not change much in the period of frame length.

Windowing After being partitioned into frames, each frame is multiplied by a window function prior to the spectral analysis to reduce the effect of discontinuity introduced by the framing process by attenuating the values of the samples at the beginning and end of each frame. Commonly used windows include Hamming and Hanning windows. If no window is used, the case can be treated as the rectangular window. Each window has its own pros and cons. Compared to rectangular window, the Hamming and Hanning windows decrease the frequency resolution of the spectral analysis while reducing the sidelobe level of the window transfer function [39].

Spectral estimation The spectral coefficients of the the speech frames are estimated using the fast Fourier transform (FFT) algorithm for MFCC. These coefficients are complex numbers containing both magnitude and phase information. For speech recognition tasks, the phase information is usually discarded and only the magnitude of the spectral coefficients are extracted. It is also common to use the power of the spectral coefficients. Besides FFT, there is another spectral estimation technique called linear predictive coding (LPC) analysis which is used to extract the LPC cepstral coefficients. One difference between the LPC spectral analysis and FFT spectral analysis is that the LPC spectrum is a parametric estimate of the smoothed spectral envelope, while the FFT spectrum tends to provide more details of the spectrum of the speech frame.

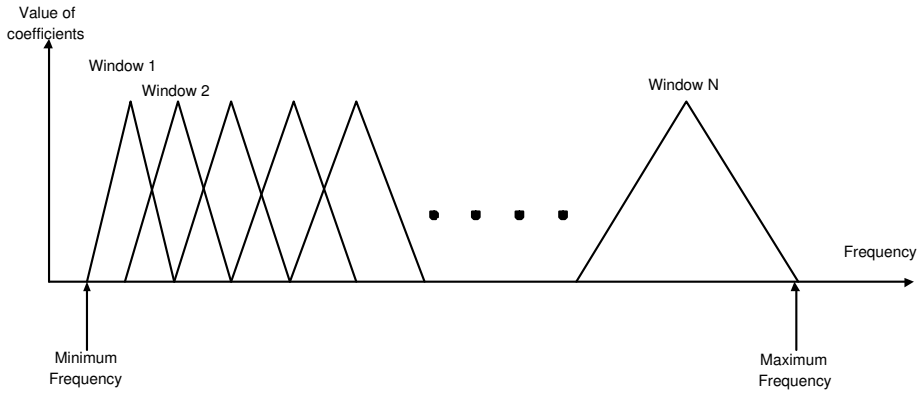

Figure 2.2: The illustration of MEL windows.

Mel flitering The spectrum of speech signal is then filtered by a group of triangle bandpass filters that simulate the characteristics of human's ear. These windows are called the Mel windows and the filtering process is called Mel filtering. The Mel filtering is to model the human auditory system that perceives sound in a nonlinear frequency binning [44]. For example, the musical pitch described in octaves and semitones is basically proportional to the logarithm of frequency. The ears analyze the spectrum of the sound in groups according to a series of overlapped critical bands. The critical bands is distributed in a way that the frequency resolution is high in low frequency region and low in high frequency region.

There are several ways to distribute the critical bands and Mel frequency scale is one of them (See Fig. 2.2). The bandwidth of the window is narrow in low frequency and gradually increases for higher frequency. The edge of the window is arrange so that it coincides with the center of the neighbor window. To decide the location of the Mel frequency of the center of the windows, the Mel frequencies for minimum and maximum linear frequency are first calculated using

$$f_{\text{Mel}} = 2595 * \log(1 + f/700) \tag{2.1}$$

where f_{Mel} is the Mel frequency for the linear frequency f. The windows are evenly distributed in the Mel frequency, and the center frequencies of the windows, when converted back to linear frequency, is not linear.

Natural logarithm While the Mel filtering approximates the nonlinear characteristics of human auditory system in frequency, the natural logarithm deals with the loudness nonlinearity. It approximates the relationship between the human's perception of the loudness and the sound intensity [46]. Besides this, it converts the multiplication relationship between parameters into addition relationship [40]. The convolutional distortions, such as the filtering effect of microphone and channel, and the multiplication in frequency domain, such as the amplification of soft sound, become simple addition after the logarithm. Hence they can be easily removed by subtracting the mean of the coefficients. This technique is called cepstral mean subtraction/normalization [78].

Discrete cosine transform The DCT is applied on the log Mel filterbank coefficients to generate the cepstral coefficients, and this process is a modified Homomorphic processing [42]. The Homomorphic processing is very useful in speech recognition, as it can separate the vocal tract shape function from the excitation signal of the speech production model. The lower order cepstral coefficients represents the smooth spectral shape or vocal tract shape, while the higher order coefficients represents the periodicity in the waveform, or the excitation information [35, 36]. Only the lower order coefficients (order$<$20) are used in speech recognition systems, hence a dimension reduction is achieved. Another benefit of DCT is that the generated cepstral coefficients are less correlated than the log Mel filterbank coefficients. Therefore, it is possible to use diagonal matrix for the covariance matrix of the Gaussian in the HMM acoustical model, and this significantly reduces the number of parameters in the acoustical model.

Log energy calculation In addition to the normal MFCC features, the energy of the speech frame is also used as a feature. The log energy, called logE, is calculated directly from the time-domain signal of a frame. Sometimes, it is replaced by C0, the 0^{th} component of the MFCC feature, which is the sum of the log Mel filterbank coefficients.

Derivatives and accelerations calculation The trend of the speech signals in time is lost in the frame-by-frame analysis. To recover the trend information, the time

derivatives (the first delta) and accelerations (second delta) are used. For speaker independent speech recognition system, the derivatives and accelerations are especially important. Although the location of the formant of the speech varies from person to person, the time trend of the formant are quite constant among different speakers. The trend information, represented by derivatives, are important for improving the robustness of the recognition.

There are several ways to approximate the first order derivative of the cepstral coefficients. For example, the derivative of coefficient $x(n)$ can be calculated as [35]

$$\dot{x}(n) \equiv \frac{d}{dt}x(n) \approx \sum_{m=-M}^{M} mx(n+m) \tag{2.2}$$

where $2M + 1$ is the number of frames considered in the evaluation. The same formula can be applied to the first order derivative to produce the second order derivative. These derived features are simply concatenated to the original cepstral features to form the final feature vector.

Normalization The normalization process ensures that all the features contribute equally. Without normalization, the feature with large dynamic range, such as the energy feature $C0$, may dominate the Euclidean distance.

2.1.2 Acoustic Models

The acoustic models are used to capture speech feature statistics in a parametric way. Currently, the dominant technique for acoustic model is the hidden Markov model (HMM) [10]. The HMM is designed to capture time varying signal's statistics and can be considered as a generalization of the Gaussian mixture model (GMM). As it is difficult to formulate a continuously time varying model, the HMM models it by a state to state transition, hence this approach can be considered as the discretization of the continuous varying case.

A general HMM consists of several inter-connected states and each state is a GMM. The model jumps from one state to another according the signal and state transition relationship between the states. The HMM used in speech recognition is a simplified

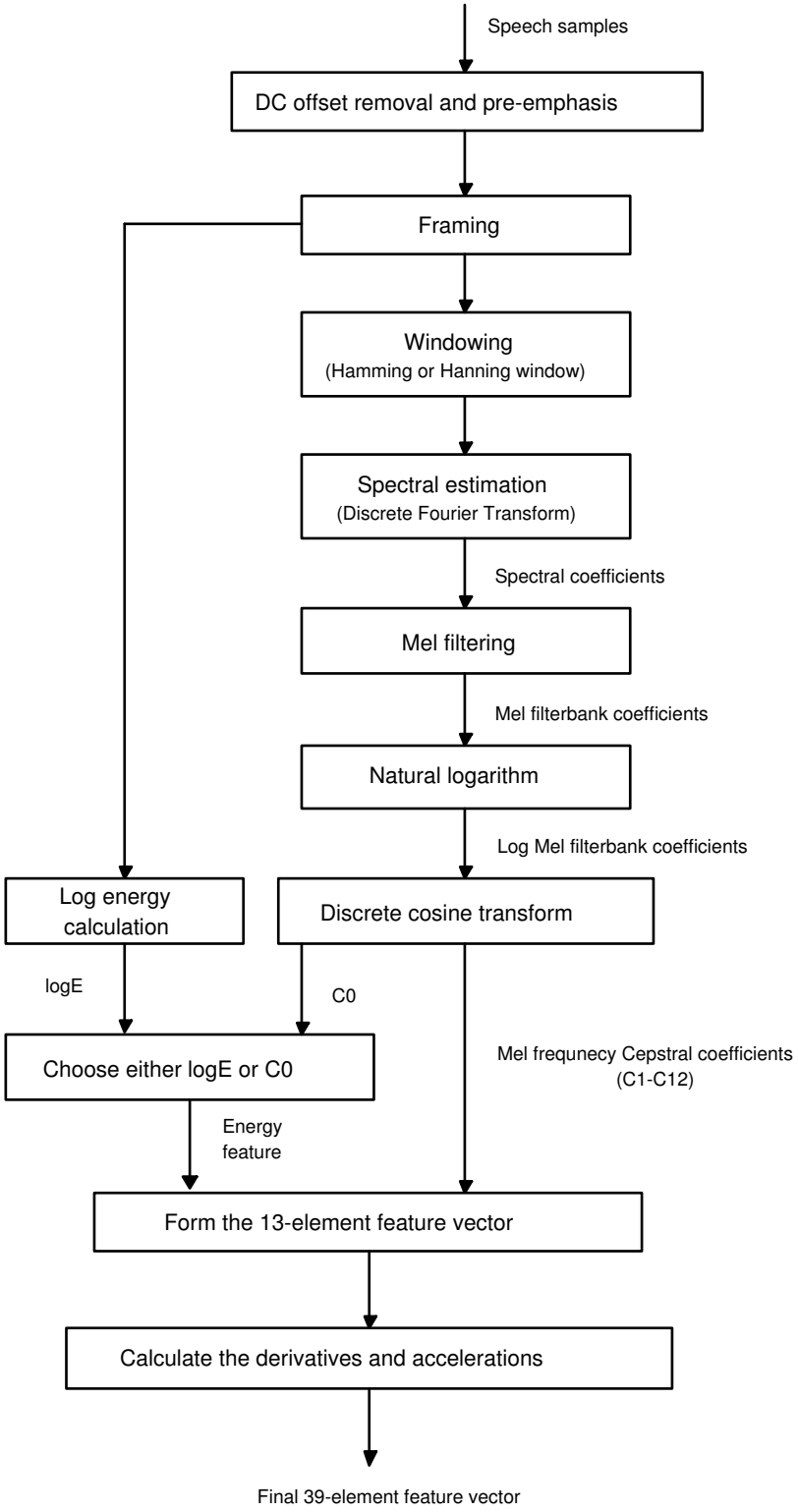

Figure 2.3: The common procedures of feature extraction (MFCC).

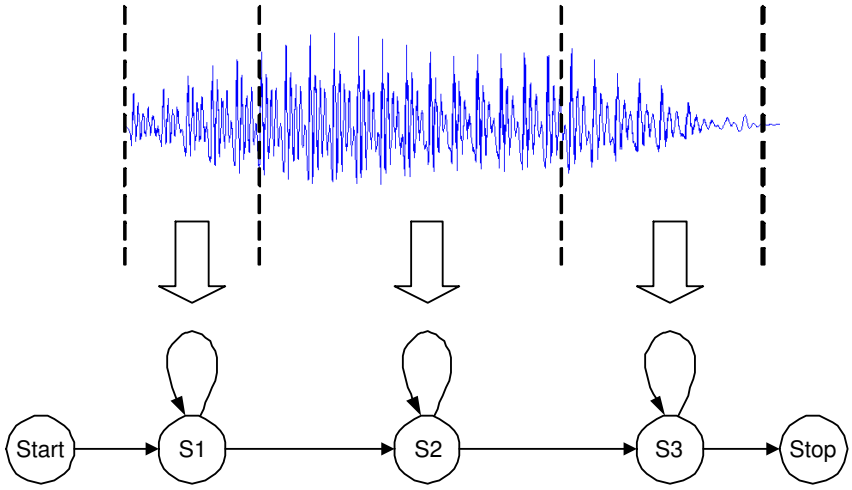

Figure 2.4: An illustration of HMM with three true states. The start and stop states are dummy ones which do not model the signal

version of the general model. It is usually a first-order HMM, i.e. the occurrence of a state is only related to its previous neighbor. In addition, the state transitions are usually constrained to be from left to right or self repetition. We denote this type of model left to right model (See Fig.2.4). In the training process of speech recognition systems, each basic speech unit, such as phoneme, is represented by a HMM. The parameters of the HMM is estimated from all the examples of the phoneme in the training data which will maximize the likelihood of the training data. As there is no closed form solution for this problem, the Expectation-Maximization (EM) algorithm [14] is used to find the solution iteratively. Once trained, these basic HMM's can be combined to form model for larger speech unit such as words and sentences, under the guide of dictionary and grammar.

2.1.3 Language Model and Word Lexicon

The word lexicon, i.e. the dictionary, contains all possible words allowed in the system. By combining the word level HMM's, all possible sentences can be generated. A large part of these possible sentences are not grammatically correct or make sense. To reduce nonsensical sentences, the language model adds constraints on the generation of the sentences to make them syntactically correct and semantically meaningful. Currently, the most popular language model is "the n-gram word grammar, where the conditional probability of each word in the sentence is a function of the previous $n-1$ words, and

the probability of a sentence (a sequence of words) is the product of these conditional probabilities" [1]. The model is trained using a large amount of text, such as newspaper, etc.

2.1.4 The Pattern Classifier and Confidence Scoring

The pattern classifier takes in all the relevant information, such as the input feature vectors, the acoustic and language model and the word lexicon to make the best guess of the underlying sentence spoken by the user. In the classification process, all the possible sentences are examined and the final result is the sentence with the largest likelihood. The likelihood of a sentence can be decomposed into two parts, the acoustic likelihood score yield by the HMM's and the language likelihood score generated by the language model. As the frames of the signal are assumed to be independent, the likelihood of the sentence can be calculated as the multiplication of the frame likelihood.

A problem of the decoding process is that the number of possible sentences grows exponentially with the length of the sentence, and for a moderate sentence length, such as ten seconds, this number can be huge. This explosive growth of possibilities is solved by pruning the least likely candidates of the recognized sentence and keeps only the N best candidates with the highest likelihood during the search[9].

In order to verify the output of the recognizer, it is necessary to have a confidence measure. If the confidence measure of the output is low, it implies that the recognized words or sentence are highly unlikely, in which case the output may be rejected.

2.2 Noise's impact on Speech Recognition

Although the word accuracy of the start-of-art ASR system is high under clean testing environments, the accuracy degrades quickly when the test speech is corrupted by noise. The work of Dautrich et. al. in 1983 [34] demonstrated the noise's impact in the isolated word recognizer (See Fig.2.5), which is also an indicator for larger tasks.

Three experiments were conducted in [34]:

(i) Training on clean speech and testing on noisy speech at different SNR levels (line: $\triangle$). The results show an increasing degradation in recognition accuracy as the SNR levels becomes lower.

(ii) Matched training and testing case across several SNR levels (line: •). The degradation in recognition accuracy obtained is much more graceful than that of the clean training case in experiment 1.

(iii) Testing on SNR=18dB speech and training on speech at various SNR levels (line: □). It is observed that the degradation in recognition accuracy is proportional to the SNR differences between the training and testing speech.

From the experimental results, we summarized two impacts of noise on speech recognition. First is the mismatch between the statistics of the training speech (represented in acoustical models) and the testing speech. The mismatch reduces the effectiveness of the clean trained acoustical models and causes the recognition accuracy to fall. Higher SNR difference between the training and testing speech causes higher degree of mismatch, and therefore result in greater degradation in recognition performance.

The second effect of noise is to reduce the distance between the basic speech units and therefore deteriorates the differentiability of the acoustic model. Many distinguishing speech information, especially those with weak energy, are distorted or lost. Hence, the noisy feature vectors from different phonemes become similar and difficult to be differentiated. This is illustrated by the matched training-testing case, where that the accuracy in low SNR levels is poorer than clean training-testing experiments (line: △, clean testing case).

Although the matched training-testing results show good robustness to noise, it is difficult to implemente as we don't have the training data exactly the same as the testing data in real world applications. In addition, the testing environment is unpredictable and only a clean database is available for training most of time. The objective of the noise robust speech recognition technique is to approach the performance of clean training-testing experiments in adverse acoustic environment where the noises are unknown and changing. In the following chapters, we will review several techniques that achieved relative success in this objective.

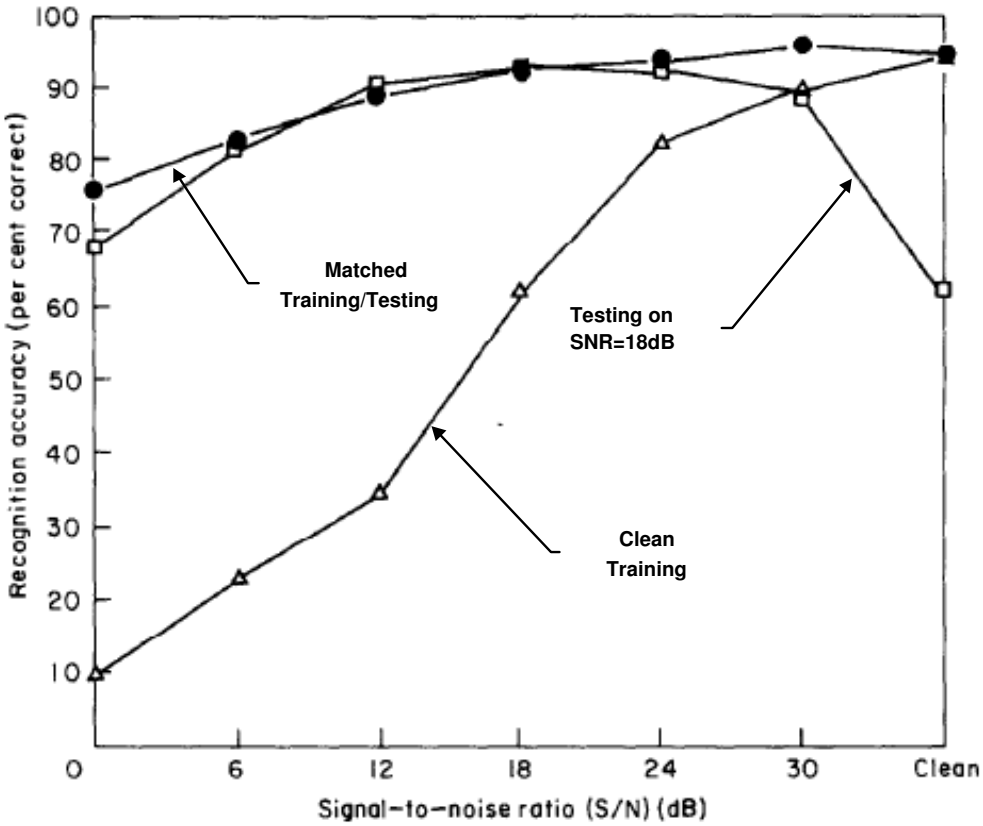

Figure 2.5: The impact of noise in speech recognition accuracy. •: training and testing have matched SNR; △: clean training and noisy testing on different SNR levels indicated by the abscissa; □: testing on SNR=18dB condition and training on different SNR levels indicated by the abscissa. (after [34, 43])

Chapter 3

Current Techniques on Noise Robust ASR

In this section, we review the existing techniques for noise robust speech recognition in three groups, namely the speech enhancement techniques in signal space, the feature compensation techniques in feature space and the the model adaptation techniques in model space (see Figure 3.1). In signal space speech enhancement techniques, the idea is to enhance the noisy signal prior to the feature extraction using enhancement techniques such as Wiener filter and spectra subtraction [17]. In feature space enhancement techniques, the noisy features are transformed to clean features through the reverse transform of D_2 aims to bring noisy feature statistics closer to the clean features to match feature to trained model. In the model space enhancement techniques, the idea is to adapt the clean trained models to better represents the noisy features.

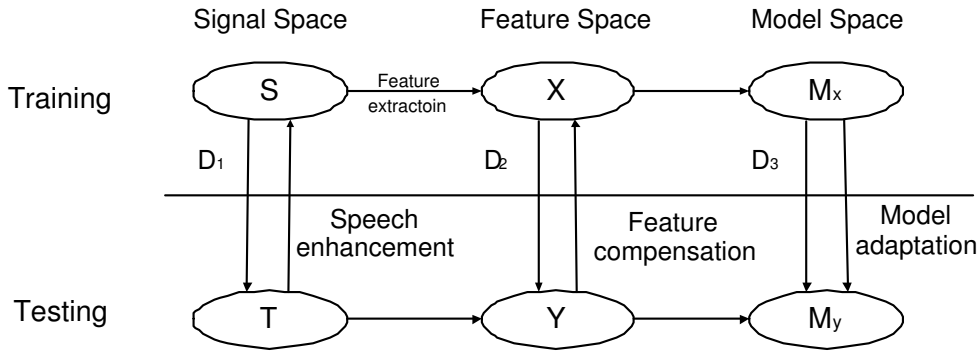

Figure 3.1: The acoustic mismatches in signal, feature and model spaces (adapted from [62]).

3.1 Preliminaries

Before discussing the various techniques, we first introduce some preliminaries, including the Bayesian estimation theory, the modeling of the prior speech statistics and the environment model that explicitly specifies how the clean speech is corrupted by environmental interferences.

3.1.1 Estimation Theory

Estimation problem is common in feature compensation techniques, such as the estimation of the clean speech, noise characteristics and channel distortion function. The accuracy of the estimation is upper bounded by several factors, such as the amount of information available and the complexity of the problem. In feature compensation context, the ability to provide a good estimate of these variables depends on three factors: the efficient use of relevant information (the prior information), the correct modeling of the transfer function between the clean and noisy speech features (the environment model) and the effectiveness of the estimation procedure (the estimation theory). We first discuss the estimation theory in this section and the other two topics in the following two sections.

There are two class of estimation theories, namely the classical estimation theory and the Bayesian estimation theory [45]. The most significant difference between the two is that in the classical estimation theory, the parameter to be estimated is treated as an unknown constant; while in the Bayesian estimation theory, it is considered as a unknown random variable. If the prior information about the parameter is available, the Bayesian estimation theory enables the use of the prior information to produce more accurate estimate than the classical estimation theory [45].

Before discussing the details of estimation theory, we first establish notations and the formulation of the problems. Let the continuous-valued observation data $\mathbf{y} = \{y_1, ..., y_N\}$ be modeled as random variables, and let its probability density function (PDF) be $p(\mathbf{y}|\Theta)$ where Θ is the parameter of the PDF. Depending on the purpose of the task, there are several estimation problems. One problem may be to estimate the unknown parameter Θ from the observation data $\mathbf{y}$, e.g., the training of the acoustical model of the speech

recognition engine. Another problem may be to estimate the clean observation from the noisy observation given the known parameter Θ, e.g. the enhancement of the noisy speech signal and the compensation of the noisy speech features problems.

Now we clarify three terms used in estimation, the *estimate*, *estimation* and *estimator*. An *estimator* is a function whose inputs are the available data and output is an estimated value of the unknown variable. The estimated value from the *estimator* is called the *estimate* of the unknown variable and the process to obtain the *estimate* is called *estimation*. E.g., to estimate the parameter Θ from observation $\mathbf{y}$, the *estimator* $f(\mathbf{y})$ takes in the observation $\mathbf{y}$ to yield an *estimate* $\hat{\Theta}$ of parameter Θ. The process to obtain $\hat{\Theta}$ is called *estimation*. In the next subsections, the two estimation theories are discussed.

3.1.1.1 Classical estimation theory

The maximum likelihood (ML) criterion is the most popular criterion in classical estimation theory. To illustrate, we assume the estimation problem is to estimate the model parameter Θ from the observation vector $\mathbf{y}$. The ML estimate of Θ is obtained by maximizing the likelihood of the observation with respect to Θ.

$$\hat{\Theta}_{ML} = \arg\max_{\Theta}\{P(\mathbf{y}|\Theta)\} \tag{3.1}$$

In speech recognition system, the HMM acoustical models are estimated using ML criterion realized by the EM algorithm.

3.1.1.2 Bayesian estimation theory

There are two popular criteria in Bayesian estimation theory, they are the minimum mean square error (MMSE) criterion and the maximum *a posteriori* (MAP) criterion. To illustrate them, we assume the estimation problem is to estimate the clean speech feature $\mathbf{x}$ from the corresponding noisy speech feature $\mathbf{y}$. The MMSE estimate $\hat{\mathbf{x}}_{MMSE}$ is defined as the estimate that minimized the mean square error between the estimate and the true value.

$$\hat{\mathbf{x}}_{MMSE} \triangleq \arg\min_{\mathbf{x}}\{E[||\hat{\mathbf{x}} - \mathbf{x}||^2]\} \tag{3.2}$$

$$= \arg\min_{\mathbf{x}}\left\{\int_{\mathbf{x}}\int_{\mathbf{y}} ||\mathbf{x} - \hat{\mathbf{x}}||^2 p(\mathbf{y}, \mathbf{x}) d\mathbf{y} d\mathbf{x}\right\} \tag{3.3}$$

where $||\cdot||^2$ represents the Euclidian distance and $p(\mathbf{y},\mathbf{x})$ is the joint probability of $\mathbf{y}$ and $\mathbf{x}$. The MMSE estimate $\hat{\mathbf{x}}_{MMSE}$ is found to be [45]

$$\begin{aligned}\hat{\mathbf{x}}_{MMSE} &= E[\mathbf{x}|\mathbf{y}] \\ &= \int_{\mathbf{x}} \mathbf{x}p(\mathbf{x}|\mathbf{y})d\mathbf{x} \\ &= \frac{\int_{\mathbf{x}} \mathbf{x}p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}}{p(\mathbf{y})} \\ &= \frac{\int_{\mathbf{x}} \mathbf{x}p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}}{\int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}}\end{aligned} \tag{3.4}$$

The MAP estimate $\hat{\mathbf{x}}_{MAP}$ is obtained by maximizing the *a posteriori* probability of the clean feature $\mathbf{x}$ given the noisy observation $\mathbf{y}$

$$\begin{aligned}\hat{\mathbf{x}}_{MAP} &\triangleq \arg\max_{\mathbf{x}}\{p(\mathbf{x}|\mathbf{y})\} \\ &= \arg\max_{\mathbf{x}}\{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\}\end{aligned} \tag{3.5}$$

Comparing the solution of the MMSE estimator (equation (3.4)) and the definition of the MAP estimator (equation (3.5)), we find that the MMSE estimate is the mean of the conditional probability density function $p(\mathbf{x}|\mathbf{y})$ and the MAP estimate is the mode of $p(\mathbf{x}|\mathbf{y})$. The evaluation of them requires the prior distribution of the clean speech $p(\mathbf{x})$ and the likelihood of the noisy speech $p(\mathbf{y}|\mathbf{x})$. The prior $p(\mathbf{x})$ represents our knowledge about the clean feature variables before any observation is seen, and the likelihood function $p(\mathbf{y}|\mathbf{x})$ contains our knowledge about $\mathbf{y}$ when we know $\mathbf{x}$ [47].

The Bayesian estimation theory provides a platform to combine the prior information and the posterior observation to make the best estimate of the unknown variables. The performance of Bayesian estimation is better than the ML estimation if useful prior information is available and properly used. A general approach to estimate the clean feature is to first define a proper format for the prior distribution $p(\mathbf{x})$ of the clean speech feature $\mathbf{x}$, and then derive the conditional probability distribution of the noisy observation $\mathbf{y}$, $p(\mathbf{y}|\mathbf{x})$, using the environment model, and finally estimate the underlying clean feature $\hat{\mathbf{x}}$ using either the MMSE and MAP criterion.

3.1.2 Modeling the Speech Distribution

The statistical distribution of the clean speech is an important information source for the enhancement of the speech and features. Although current statistical models for speech distribution, such as the Gaussian mixture model (GMM) and hidden Markov model (HMM), models the speech distribution effective, they generally ignores the inter-frame relationship of speech. GMM assumes that the speech feature vectors are independent from each other and are generated from several multivariate random processes and one Gaussian is used to represent one process. Given enough number of mixtures, the GMM can approximate any complex distribution function. However, more mixtures implies more parameters needed to be trained, which requires more training data. The number of mixture is a compromise between the model complexity and availability of enough training data, and can be experimentally determined. For a random vector $\mathbf{x}$, the GMM distribution function is defined as

$$p(\mathbf{x}) = \sum_{k=1}^{K} P(k)\mathcal{N}(\mathbf{x}, \mu_k, \Sigma_k) \tag{3.6}$$

where K is the number of Gaussian mixtures, $P(k)$ is the prior weight of the mixtures and $\mathcal{N}(\mathbf{x}, \mu_k, \Sigma_k)$ is the distribution function of the k^{th} mixture. μ_k and Σ_k are the mean vector and covariance matrix of a Gaussian mixture respectively. The parameters of the model can be trained using the EM algorithm [14]. The GMM can model the non-stationary speech process as it uses different Gaussian for different underlying processes. However, the trend information of speech in time cannot be captured by this model. The HMM is an extension of GMM consists of multiple states and the state-transition probability matrix to capture the time information, However, its independent assumption for neighboring frames limits the full representation of the time information. A simple method to incorporate the time information is the use of the derivative features in speech recognition systems. In addition, several variants of HMM are proposed to compensate the independence assumption, including trended HMM [11], segmental HMM [12] and linear predictive HMM [13]. In the trended HMM, the Gaussian means in each HMM state are characterized by time-varying polynomial trend functions of the state sojourn time. The segmental HMM captures the probability of a segment of observation frames

rather than the probability of one observation frame in the classical HMM. In the linear predictive HMM, the observation vector of time t is linearly predicted from the previous P frames, where P is the order of the linear prediction.

A statistical way of modeling the inter-frame relationship was proposed by Raj [77, 65] for missing feature reconstruction. It assumes that the feature vectors are generated from a single stationary process and the cross-covariance of neighboring frames are estimated. The single process assumption limited the ability of the model to capture the statistics of non-stationary speech signal. In chapter 5, we proposed a vector-autoregressive model for modeling the inter-frame relation of the speech feature vectors.

3.1.3 Environment Model

The environment model specifies how the noisy speech signal is generated from the clean speech signal, the noises and channel distortions. Generally, the techniques using an environment model are called model-based techniques while others without using an environment model is called data-driven techniques. In this section, we will introduce the commonly used environment model in the noise robust speech recognition field.

According to the way the noise corrupts the speech signal, the noise can be classified into three categories, they are the additive noise, multiplicative noise and convolutionary noise. In this report, we only examine the additive noise and convolutionary noise. Example of additive noises are background noise, traffic noise, etc., and convolutionary noise may be transmission channel distortions, microphone filtering, etc. The environment model we are going to discuss is shown in Fig.3.2, where the clean speech signal is distorted by the channel first, then corrupted by the additive noise further.

We derive the mathematical representation of the environment model now. Let $x(n)$, $n(n)$ and $y(n)$ represent the digital clean speech, additive noise and degraded speech in time domain respectively, and let $h(n)$ represent the channel impulse response. The environment model in time domain is:

$$y(n) = x(n) * h(n) + n(n) \tag{3.7}$$

where $*$ represents convolution. Applying the discrete Fourier transform, the model in frequency domain for a single frame is

$$Y(k) = X(k)H(k) + N(k) \tag{3.8}$$

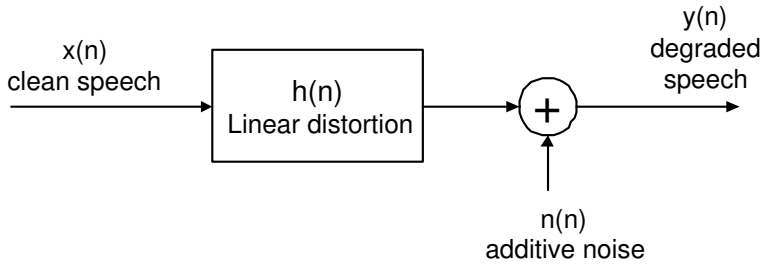

Figure 3.2: The environment model with additive noise and linear distortion.

where $k = 1, ..., K$ is the number of Fourier coefficients, $Y(k)$, $X(k)$, $N(k)$ and $H(k)$ are the Fourier transform coefficients of

$y(n)$, $x(n)$, $n(n)$ and $h(n)$ in current frame respectively. The convolution in equation (3.7) becomes multiplication in equation (3.8). The power spectral density of noisy speech signal is found using:

$$\begin{aligned} |Y(k)|^2 &= Y(k)Y^*(k) \\ &= (X(k)H(k) + N(k))(X^*(k)H^*(k) + N^*(k)) \\ &= X(k)X^*(k)H(k)H^*(k) + X(k)H(k)N^*(k) + X(k)^*H^*(k)N(k) + N(k)N^*(k) \\ &= |X(k)|^2|H(k)|^2 + |N(k)|^2 + X(k)H(k)N^*(k) + (X(k)H(k)N^*(k))^* \end{aligned} \quad (3.9)$$

By using the property that the product of two complex number equals to the product of their product and the cosine of the angle between them, i.e., $ab = |a||b|cos(\alpha)$, the equation (3.9) can be rewritten as

$$|Y(k)|^2 = |X(k)|^2|H(k)|^2 + |N(k)|^2 + 2|X(k)||H(k)||N(k)|\cos\theta_k \quad (3.10)$$

$$\cos\theta_k = \frac{X(k)H(k)N^*(k)}{|X(k)||H(k)||N(k)|} = \frac{X^*(k)H^*(k)N(k)}{|X(k)||H(k)||N(k)|} \quad (3.11)$$

where θ_k denotes the random angle between the two complex variables $N^*(k)$ and $X(k)H(k)$, and it contains the phase information. The environment model has two versions depends on whether they include the phase term or not. We will introduce them separately in the following subsections.

3.1.3.1 The phase-insensitive model

If the phase term is ignored in equation (3.9), the power spectral density of $Y(k)$ is simplified to

$$|Y(k)|^2 \approx |X(k)|^2|H(k)|^2 + |N(k)|^2 \quad (3.12)$$

Note that this simplification is due to the fact that $E[2|X(k)||H(k)||N(k)|\cos\theta_k] = 0$, because the noise and speech are assumed to be independent and that the noise has zero mean. This simplification causes the lost of phase information as the equation (3.12) only holds in expected sense. After obtaining the power spectrum, the L Mel filterbanks coefficients are computed (this part of derivation follows that in [56]). Let $w_k^l, k = 1, ..., K/2$ denote the Mel window weights for the l^{th} filterbank, where $\sum_{k=1}^{K/2} w_k^l = 1$. During the evaluation, half of the linear power spectrum are discarded due to its symmetrical structure. The calculation of the filterbank coefficients is as follows

$$\begin{aligned} |\tilde{Y}^l|^2 &= \sum_{k=1}^{K/2} w_k^l |Y(k)|^2, \quad l = 1, ..., L \\ &= \sum_{k=1}^{K/2} w_k^l (|X(k)|^2 |H(k)|^2 + |N(k)|^2) \\ &= |\tilde{X}^l|^2 |\tilde{H}^l|^2 + |\tilde{N}^l|^2 \end{aligned} \tag{3.13}$$

where

$$|\tilde{X}^l|^2 = \sum_{k=1}^{K/2} w_k^l |X(k)|^2 \tag{3.14}$$

$$|\tilde{N}^l|^2 = \sum_{k=1}^{K/2} w_k^l |N(k)|^2 \tag{3.15}$$

$$|\tilde{H}^l|^2 = \frac{\sum_{k=1}^{K/2} w_k^l (|X(k)|^2 |H(k)|^2}{|\tilde{X}^l|^2} \tag{3.16}$$

Applying natural logarithm and DCT on the filterbank coefficients yields

$$\text{DCT}(\ln |\tilde{Y}^l|^2) = \text{DCT}(\ln(|\tilde{X}^l|^2 |\tilde{H}^l|^2 + |\tilde{N}^l|^2)), \quad l = 1, ..., L \tag{3.17}$$

where DCT represents the discrete cosine transform. To represent equation (3.17) in a simpler manner, the following definitions are introduced

$$\mathbf{x} = \text{DCT}(\ln |\tilde{X}^l|^2) \tag{3.18}$$

$$\mathbf{n} = \text{DCT}(\ln |\tilde{N}^l|^2) \tag{3.19}$$

$$\mathbf{y} = \text{DCT}(\ln |\tilde{Y}^l|^2) \tag{3.20}$$

$$\mathbf{h} = \text{DCT}(\ln |\tilde{H}^l|^2) \tag{3.21}$$

Equation (3.17) can be rewritten in a simpler manner after following algebra transformation

$$\begin{aligned}
\mathbf{y} &= \text{DCT}\left\{\ln\left[|\tilde{X}^l|^2|\tilde{H}^l|^2\left(1+\frac{|\tilde{N}^l|^2}{|\tilde{X}^l|^2|\tilde{H}^l|^2}\right)\right]\right\} \\
&= \text{DCT}\{\ln|\tilde{X}^l|^2\} + \text{DCT}\{\ln|\tilde{H}^l|^2\} + \text{DCT}\left\{\ln\left[\left(1+\frac{|\tilde{N}^l|^2}{|\tilde{X}^l|^2|\tilde{H}^l|^2}\right)\right]\right\} \\
&= \mathbf{x}+\mathbf{h}+\text{DCT}\{\ln(1+\exp(\text{IDCT}[\mathbf{n}-\mathbf{h}-\mathbf{x}]))\} \qquad (3.22)
\end{aligned}$$

3.1.3.2 The phase-sensitive model

If the phase term is included in equation (3.9), there will be one phase-related term in the Mel filterbank representation of noisy speech signal (this part of derivation follows that in [56]). The equation (3.17) is expanded as

$$\ln|\tilde{Y}^l|^2 = \ln(|\tilde{X}^l|^2|\tilde{H}^l|^2 + |\tilde{N}^l|^2 + 2\alpha^l|\tilde{X}^l||\tilde{H}^l||\tilde{N}^l|), \quad l = 1, ..., L \qquad (3.23)$$

where the phase term α^l is defined as

$$\alpha^l \equiv \frac{\sum_{k=1}^{K/2} w_k^l 2|X(k)||H(k)||N(k)|\cos\theta_k}{|\tilde{X}^l||\tilde{H}^l||\tilde{N}^l|} \qquad (3.24)$$

After applying the natural logarithm and DCT, and follow the same definition as in equation (3.18-3.21), the model in cepstral domain is represented as:

$$\mathbf{y} = \mathbf{x}+\mathbf{h}+\log[1+\exp(\mathbf{n}-\mathbf{x}-\mathbf{h})+2\alpha\bullet\exp(\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2})] \qquad (3.25)$$

$$\alpha = \frac{\exp(\mathbf{y}-\mathbf{x}-\mathbf{h})-\exp(\mathbf{n}-\mathbf{x}-\mathbf{h})-1}{2\exp(\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2})} \qquad (3.26)$$

where $\alpha = [\alpha^1, \alpha^2, ..., \alpha^l]^T$ is a vector and $\bullet$ denotes the element-wise multiplication. This model is adopted in [56].

3.2 Speech Enhancement Techniques

Speech enhancement techniques are originally designed for enhancing noise-corrupted speech signals for human listening. The criteria for speech enhancement are quite different from that for feature compensation. Speech enhancement techniques usually aim at

improving the quality and intelligibility of speech signals, while the feature compensation techniques aim at reducing the mismatch between the noisy speech feature and the clean trained acoustic model. Despite this difference, it is reasonable to assume that the better the quality and intelligibility of the speech signal, the better the the speech recognition accuracy can be obtained. Therefore, many speech enhancement techniques have been modified and applied in the feature extraction process of the speech recognition system with some success. In the following sections, we will review several popular speech enhancement techniques, such as spectral subtraction [17]-[20], the Ephraim's MMSE STSA estimator [21]-[23] and the signal subspace-based techiques [28]-[33].

3.2.1 Spectral Subtraction

Spectral subtraction estimates the clean speech spectrum by subtracting the estimated additive noise spectrum from the noisy speech spectrum [17]. The noise is usually assumed to be stationary, and its power spectrum is estimated during speech absent frames. The phase information is discarded, so the noisy speech is assumed to be the sum of clean speech and noise in terms of power spectral density, which is true only in expected sense. Due to the variance of the noise, the estimated clean spectral coefficients may be negative. In such cases, its values are set to zero. This irregularity produces illegitimate spectral vectors called "musical noise" phenomenon, which is unacceptable for human listening and degrades speech recognition performance. Berouti et al. [18] proposed to use a spectral floor to reduce the "musical noise". The estimation of the clean spectrum is as follows

$$|\hat{X}(k)| = \begin{cases} \sqrt{|Y(k)|^2 - \alpha\overline{|N(k)|^2}}, & |Y(k)|^2 - \alpha\overline{|N(k)|^2} > (\beta|N(k)|)^2; \\ \beta|N(k)|, & \text{otherwise.} \end{cases} \tag{3.27}$$

where $|\hat{X}(k)|$, $|Y(k)|$ and $|N(k)|$ represents the spectral amplitude of the estimated clean signal, the noisy signal and the noise, respectively. $\overline{(\cdot)}$ is the time averaging function, α is a SNR dependent parameter for over subtraction and β is a experimentally determined spectral floor parameter. α and β enable a tradeoff between the "musical noise" and residual noise level. The larger the α, the smaller is the residual wide-band noise but the larger the speech distortion, and vice versa. Some researchers proposed to apply the

masking properties of ear to determination of the amount of subtraction [19]. The basic idea is that for those noise that can not be heard by human's ear, it is not subtracted, so there are smaller amount of subtraction and therefore smaller degree of distortion.

Despite its applications in enhancing speech for hearing, spectral subtraction have also been used to preprocess noisy speeches in feature extraction process of speech recognition systems [3]. The limitation of spectral subtraction is that it only performs well when noise is stationary and the SNR is high.

3.2.2 Ephraim's Estimator

Ephraim proposed an estimator of the short-time spectral amplitude (STSA) of the speech in the MMSE sense [21, 22]. Ephraim's estimator uses a Gaussian model for the distribution of the spectral components. Specifically, the phase and amplitude of the spectral components of the clean speech signal and noise are assumed to be independent Gaussian variables. Using statistical rules, it is found that the distribution of the spectral component of the noisy speech follows the Rayleigh's distribution. The MMSE estimate of the clean spectral amplitude is derived based on these models, and the resulting solution of the estimator is a complex function of the *a priori* SNR and the *a posteriori* SNR. The *a priori* SNR is crucial for overall performance, and can be estimated using either ML estimation or a "decision-directed" method. The *a posteriori* SNR can be interpreted as the instantaneous SNR of the current frame.

As the *a priori* SNR is important for the estimation of clean speech, several methods are proposed to improve its estimate. Many research were focused on the average weighting parameter α of the decision directed method, which controls the speech of adaptation. Soon and Koh [24] proposed to adapt the α by estimating it from the changing speed of the frame energy. This idea is further extended by Hansen et al.[25] by making α frequency dependent and applying the MMSE criterion. Besides the improvements on the estimation of α, to incorporate more information, Israel [26] proposed a noncausal *a priori* SNR estimator that employs both the previous and future spectral measurements to better estimate the *a priori* SNR. Another approach by Hu and Loizou [27] reduces the variance of the *a priori* SNR estimate indirectly by reducing the variance of the estimated noisy signal power spectrum using a new spectrum estimator.

Ephraim's estimator also incorporates a signal absence probability (SAP) first introduced by McAulay and Malpass [20]. The SAP accounts for the probability of the absence of the speech signal. During the enhancement process, the SAP of every frequency bin is calculated based on the noisy speech signal, and then the STSA estimator of the clean speech is formed as

$$\hat{X}_k = P(H_k^1|Y(k))E[X(k)|Y(k), H_k^1] \tag{3.28}$$

where H_k^1 is the hypothesis that the signal is present in the kth spectral component; $P(H_k^1|Y(k))$ is the probability of H_k^1 estimated from $Y(k)$, $E[X(k)|Y(k), H_k^1]$ is the optimal MMSE STSA estimator when SAP is not considered. Experiments shows that the incorporation of SAP into the MMSE STSA estimator further reduces the residual noise [21].

Later, this MMSE estimator of STSA is extended to log spectral domain [22] to simulate the nonlinear compression of the human auditory system. It was reported that the log MMSE STSA estimator yielded better performance than the original estimator in [21].

3.2.3 Subspace-based Techniques

The signal subspace approach is motivated by the fact that noisy speech signal can usually be decomposed into two subspaces: the signal plus noise subspace and the noise only subspace. During the enhancement process, the noise only subspace can be first nulled and the clean speech signal can be estimated from the signal plus noise subspace.

There are two methods to decompose the noisy signal into the two subspaces, namely the Singular Value Decomposition (SVD) method and the Karhunen-Loève transform (KLT). In the SVD-based method proposed by Dendrinos *et al* [28], the clean signal is reconstructed from the singular vectors corresponding to the largest singular values. It is believed that the singular vectors corresponding to the largest singular values contain the speech information, while the singular vectors corresponding to the smallest singular values contains the noise information. This approach provides large SNR gains for speech corrupted by white noise. In the Quotient SVD-based approach proposed by Jensen *et al* [29], the previous approach is extended to suppress colored noise. However, QSVD was

found to be computationally expensive and provided no method for shaping or controlling the residual noise.

Many approaches also use the KLT to decompose the noisy signal. In Ephraim and Van Trees's method [30], the estimator minimized the speech distortion subject to a given residual noise level constraint. In this way, a mechanism is provided to adjust the tradeoff between the signal distortion and the residual noise level. Huang and Zhao [31] extended the method of Ephraim and Van Trees by proposing an energy-constrained signal subspace method (ECSS). The idea was to match the short-time energy of the enhanced speech signal to the unbiased estimated of the clean speech. They declared that this method recovered the low-energy segments in continuous speech effectively. Rezayee and Gazor [32] proposed an algorithm to reduce colored noise by diagonalizing the noise correlation matrix using the estimated eigenvalues of the clean speech and nulling any off-diagonal elements. Mittal and Phamdo [33] extended Ephraim and Van Trees's method to colored noise by providing proper noise shaping for colored noise without pre-whitening.

One important assumption of signal subspace approach is that the largest singular values or eigenvalues are from speech and the smallest values are from noise. However, in very noisy cases such as SNR=-5dB, the noise power may be higher than the signal power and this assumption may not hold any more.

3.3 Feature Compensation Techniques

Feature compensation techniques compensate the speech features in the feature domain. In this section, we surveyed the normalization techniques [78], the model based feature estimation methods [73, 55, 56, 57] and the missing feature techniques [65, 66, 68].

3.3.1 Normalization Techniques

There are several normalization techniques that are both simple and effective, such as the CMN (also known as cepstral mean subtraction (CMS)), cepstral variance normalization (CVN), vocal tract length normalization (VTLN) and histogram normalization and rotation. In Molau's PhD thesis [78], these normalization techniques are summarized nicely.

CMN is the simplest, yet very effective way of removing convolutional noises, such as linear distortion caused by different recording devices and communication channels. Due to the natural logarithm in the feature extraction process, the linear filtering usually results in a constant offset in filterbank or cepstral domain and hence can be subtracted from the signal in a sentence by sentence basis. The basic CMN [49] estimates the sample mean vector of the cepstral vectors of a sentence and then subtract this mean vector from every cepstral vector of the sentence. Later, an augmented cepstral normalization method [50] estimated the mean vector for the silence and speech segments of the sentence separately and achieved better results. Instead of using a hard decision on whether a frame is silence or speech, one improvement suggests the use of the *a posteriori* probability of the frame of being silence $p(n)$, and the final mean vector is the weighted sum of the silence mean and speech mean, with the weights be $p(n)$ and $1 - p(n)$ respectively. The advantage of CMN techniques is their simplicity, low computational cost and easy to be implemented. However, their performance is limited as they use very few prior information about the speech and noise.

Another normalization techniques is the CVN method which normalizes the variances of the cepstral coefficients to unit variance. It is found that CVN is useful for databases with large acoustic variation [78].

VTLN is a technique to normalize the difference in human's vocal tract length. Human's vocal tract length is different for different person, especially between male, female and child. As the formant of the voiced sounds is inversely proportional to the vocal tract length, different vocal tract length results in shift of the location of the formant. VTLN tries to shift the mean formant frequency to a reference value through transformation. The reference formant frequency is the average formant frequency of all speakers, and the transformation can be applied in time, either frequency or cepstral domain. VTLN techniques usually yield a relative error rate reduction of 10% [48].

Unlike CMN and CVN, which only normalize the mean and variance of the speech distributions, the histogram normalization aims at matching the speech distribution of the testing data with that of the training data. Histogram normalization assumes that the global statistics of speech is independent of topic and that the variation corrected by histogram normalization is uncorrelated among different dimensions. The distribution

of both training and testing data are brought to match the reference distribution in a dimension by dimension basis, which is defined as the global distribution of both testing and training data. Despite its conceptual simplicity, histogram normalization is quite effective in improving the recognition accuracy. Molau [78] reported that it can be applied in various stages of feature extraction and performs best in log Mel filterbank domain. This technique is more beneficial if applied to problems with a larger acoustic mismatch between testing and training data.

3.3.2 CMU's Techniques

3.3.2.1 The CDCN algorithm

Acero [73] of the Robust Speech Recognition Group of CMU proposed the codeword dependent cesptral normalization (CDCN) technique to handle the noise and linear distortion at the same time. It adopted the phase-insensitive environment model in equation (3.22) and rearranged the terms as:

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{h}) \tag{3.29}$$

$$\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{h}) = \mathrm{DCT}\{\ln(1 + \exp(\mathrm{DFT}[\mathbf{n} - \mathbf{h} - \mathbf{x}]))\} \tag{3.30}$$

and assumes that the noise is stationary and the linear distortion is constant during the observation period, and that the frames are statistically independent. The method's objective is to find the most likely spoken string S, environmental parameters $\mathbf{n}$ and $\mathbf{h}$, from the N observation frames $Z = \mathbf{z}_0, ..., \mathbf{z}_{N-1}$. From the Bayesian classification theory, this equals to maximizing the *a posteriori* probability:

$$p(S, \mathbf{n}, \mathbf{h}|Z) = \frac{p(S, Z|\mathbf{n}, \mathbf{h})p(\mathbf{n}, \mathbf{h})}{p(Z)} \tag{3.31}$$

If there is no *a priori* knowledge about $\mathbf{n}$ and $\mathbf{h}$, and because $p(Z)$ is a constant for a specific Z, the above problem is equivalent to maximizing the likelihood function:

$$\begin{aligned} p(S, Z|\mathbf{n}, \mathbf{h}) &= p(S|\mathbf{n}, \mathbf{h})p(Z|S, \mathbf{n}, \mathbf{h}) \\ &= p(S) \prod_{i=0}^{N-1} p(\mathbf{z}_i|S, \mathbf{n}, \mathbf{h}) \end{aligned} \tag{3.32}$$

since $p(S|\mathbf{n},\mathbf{h})$ does not depends on $\mathbf{n}$ and $\mathbf{h}$ and the frames are assumed to be independent.

Gaussian mixture model is used to model the probability density function of the clean speech features:

$$p(\mathbf{x}) = \sum_{k=0}^{K-1} P(k)\mathcal{N}_x(\mu_k, \Sigma_k) \tag{3.33}$$

where K is the number of mixtures, $P(k)$ is the prior weight of the mixtures and $\mathcal{N}_x(\mu_k, \Sigma_k)$ represents the distribution function of a single mixture with mean μ_k and covariance Σ_k. To simplify the problem, the correction vector $\mathbf{r}(\mathbf{x},\mathbf{n},\mathbf{h})$ is assumed to be a constant vector for each mixture and can be calculated as $\mathbf{r}(\mu_k,\mathbf{n},\mathbf{h})$, denoted as $\mathbf{r}_k$. As $\mathbf{h}$ is also assumed to be a constant vector, according to equation (3.29), the distribution of the noisy cepstral vector $\mathbf{y}$ is also a Gaussian $\mathcal{N}_y(\mathbf{h}+\mathbf{r}_k+\mu_k, \Sigma_k)$. Due to the limitation in the length of the frame used for spectral analysis, the observed cepstral vector $\mathbf{z}$ is an estimate of the true cepstral vector of the underlying random process and is assumed to be a random vector with a Gaussian distribution $\mathcal{N}_z(\mathbf{y}, \Gamma)$, centered at the true noisy cepstral vector $\mathbf{y}$ with a covariance matrix Γ. After some derivations, it is proved that the conditional probability distribution of the noisy observed cepstral vectors $\mathbf{z}_i$ also follows a GMM distribution

$$p(\mathbf{z}_i|S,\mathbf{n},\mathbf{h}) = \sum_{k=0}^{K-1} P_i(k|S)\mathcal{N}_z(\mathbf{h}+\mathbf{r}_k+\mu_k, \Gamma+\Sigma_k) \tag{3.34}$$

where $P_i(k|S)$ is the conditional probability of cluster k in the i^{th} frame. This is easy to understand, as the $\mathbf{h}+\mathbf{r}_k+\mu_k$ can be seen as the expected value of $\mathbf{y}$ (see equation (3.29)), and the variation of $\mathbf{z}$ is the sum of the variation from spectral analysis and the variation caused by noise and distortion. Substitute equation (3.34) into (3.32), we get:

$$p(S,Z|\mathbf{n},\mathbf{h}) = p(S)\prod_{i=0}^{N-1}\sum_{k=0}^{K-1} P_i(k|S)\mathcal{N}_z(\mathbf{h}+\mathbf{r}_k+\mu_k, \Gamma+\Sigma_k) \tag{3.35}$$

The optimization problem in equation (3.35) requires the combination of the GMM of $\mathbf{z}$, the HMM acoustical model $P_i(k|S)$ and the grammar information $p(S)$. To simplify the problem, a suboptimal solution is obtained solely based on the GMM of $\mathbf{z}$, which is

called the CDCN's solution. This distribution of $\mathbf{z}$ is simplified as

$$p(Z|\mathbf{n},\mathbf{h}) = \prod_{i=0}^{N-1}\sum_{k=0}^{K-1} P_i(k)\mathcal{N}_z(\mathbf{h}+\mathbf{r}_k+\mu_k, \Gamma+\Sigma_k) \tag{3.36}$$

The CDCN algorithm first estimates the environmental parameters $\mathbf{n}$ and $\mathbf{h}$ for a sentence by maximizing the likelihood function $P(Z|\mathbf{n},\mathbf{h})$, i.e., find the $\mathbf{n}$ and $\mathbf{h}$ that maximize the probability of Z. As there is no close form solution, the expectation-maximization (EM) algorithm is used to for the estimate. After the $\mathbf{n}$ and $\mathbf{h}$ are obtained, the minimum mean square error (MMSE) estimation of the clean cepstral vectors are estimated on a frame-by-frame basis

$$\begin{aligned}\hat{\mathbf{x}} &= \mathbf{z} - \hat{\mathbf{h}} - \sum_{k=1}^{K} P(k|\mathbf{z})\hat{\mathbf{r}}_k \\ \hat{\mathbf{r}}_k^j &= \mathrm{DCT}\{\ln(1+\exp(\mathrm{IDCT}[\hat{\mathbf{n}}-\hat{\mathbf{h}}-\mu_k])), j=1,...,J\}\end{aligned} \tag{3.37}$$

where $\hat{\mathbf{h}}$ and $\hat{\mathbf{n}}$ are the estimate of the channel distortion and noise, $\hat{\mathbf{r}}_k^j$ is the estimate of the j^{th} element of the correction vector from the k^{th} mixture and J is the dimension of the cepstral vector. $P(k|\mathbf{z})$ is the a posteriori probability of mixture k given the noisy observation $\mathbf{z}$.

The CDCN algorithm learns the *a priori* clean speech information in the form of GMM and analytically estimates the environment interferences to compensates the cepstral vectors in a soft decision vector quantization (VQ) fashion. Therefore, the finer the space of $\mathbf{x}$ is partitioned, the better is the approximation of VQ. Therefore, CDCN can be seen as an extension to CMN, i.e. instead of using one global correction vector for all the frames in one utterance, different correction vectors are used for every frame by computing the weighted sum of a group of predefined correction vectors. CDCN does not require the a priori knowledge of the testing environment. However, when the noise is not stationary and the linear distortion is time-varying, CDCN is expected to perform poorly.

3.3.2.2 RATZ

Later also in CMU, Moreno [76] examined the effect of noise on the statistical distribution of the log Mel spectral vectors both analytically and using the Monte-Carlo simulations.

No environment model is used and the relationship between the distribution functions of the clean and noisy speech signal are found using data-driven approach. Moreno observed that in the log Mel filterbank domain, the effect of the noise on the distribution of the speech signal is that the mean is shifted and the variance is either decreased or increased depending on the SNR. If we assume the distribution functions to be Gaussians, the distribution of the noisy speech can be approximated by adding a correcting term to the mean and variances of the clean speech distribution. This observation is also applicable to the cepstral features distribution. Let the distribution of the cepstral vectors of the clean and noisy speech be GMM with the same number of mixtures

$$p(\mathbf{x}) = \sum_{k=0}^{K-1} P(k)\mathcal{N}_x(\mu_x^k, \Sigma_x^k) \tag{3.38}$$

$$p(\mathbf{y}) = \sum_{k=0}^{K-1} P(k)\mathcal{N}_y(\mu_y^k, \Sigma_y^k) \tag{3.39}$$

The noisy distribution function can be approximated by adding correction terms to the clean mean and covariance parameters

$$\mu_y^k = \mu_x^k + \mathbf{r}^k, k = 0, ..., K-1 \tag{3.40}$$

$$\Sigma_y^k = \Sigma_x^k + \mathbf{R}^k, k = 0, ..., K-1 \tag{3.41}$$

Based on this finding, a data-driven compensation approach is proposed. The correction items $\mathbf{r}^k$ and $\mathbf{R}^k$ are first estimated from testing data, and then used for either feature vector compensation or HMM model adaptation. The feature vector compensation algorithm is called the multivaRiate gAussian-based cepsTral normaliZation (RATZ) and the model adaptation algorithm is called STAtistical Reestimation (STAR). The correction terms are estimated based on the maximization of the likelihood for the noisy observation

$$\begin{aligned}\phi &= \{\mathbf{r}^1, ..., \mathbf{r}^K; \mathbf{R}^1, ..., \mathbf{R}^K\} \\ &= \arg\max_{\phi} \prod_{n=1}^{N} p(\mathbf{y}_n) \\ &= \arg\max_{\phi} \prod_{n=1}^{N} \sum_{k=0}^{K-1} P_n(k)\mathcal{N}_y(\mathbf{y}_n, \mu_x^k + \mathbf{r}^k, \Sigma_x^k + \mathbf{R}^k)\end{aligned} \tag{3.42}$$

As there is no closed-form solution for the correction terms, the EM algorithm is applied again. After the $\mathbf{r}^k$ and $\mathbf{R}^k$ are obtained, the RATZ estimates the clean cepstral vector using the MMSE criteria.

3.3.3 Microsoft's Techniques

A group of researchers in Microsoft, Deng, Droppo and Acero [55, 56, 57] have developed several feature compensation and model adaptation techniques based on a phase-sensitive speech distortion model. In this section, we first introduce the distortion, and then discuss on their proposed techniques.

3.3.3.1 Derivation of noisy speech distribution

The phase sensitive environment model of equation (3.43) is used here.

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \log\left[1 + \exp(\mathbf{n} - \mathbf{x} - \mathbf{h}) + 2\alpha \bullet \exp(\frac{\mathbf{n} - \mathbf{x} - \mathbf{h}}{2})\right] \tag{3.43}$$

$$\alpha = \frac{\exp(\mathbf{y} - \mathbf{x} - \mathbf{h}) - \exp(\mathbf{n} - \mathbf{x} - \mathbf{h}) - 1}{2\exp(\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2})} \tag{3.44}$$

where α is defined to be a sum of k independent, zero-mean random variables non-uniformly distributed over (-1, 1) (see equation (3.24)), and k equals to the number of DFT bin included in the Mel window. Therefore, the elements of α are assumed to have Gaussian distributions since the central limit theorem states that the distribution of the sum of a large number of independent random variables approaches the Gaussian distribution under certain conditions. Furthermore, the elements of α are assumed to be zero-mean and independent, so the probability distribution of α is a Gaussian

$$p(\alpha) = \mathcal{N}(\alpha; \mathbf{0}, \Sigma_\alpha) \tag{3.45}$$

where Σ_α is the diagonal covariance matrix.

The conditional probability $p_y(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h})$ can then be found using the well-known result from probability theory on determining the PDF for functions of random variables.

$$p_y(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h}) = |J_\alpha(\mathbf{y})| p_\alpha(\alpha|\mathbf{x}, \mathbf{n}, \mathbf{h}) \tag{3.46}$$

where $J_\alpha(\mathbf{y}) = 1/(\partial\mathbf{y}/\partial\alpha)$ is the Jacobian of the nonlinear transformation and $|\cdot|$ represents the determinant of a matrix. After several steps of derivation [56], it can be shown that

$$p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h}) = \frac{1}{2}|\text{diag}(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2})|\mathcal{N}(\alpha(\mathbf{x},\mathbf{n},\mathbf{h},\mathbf{y}),\mathbf{0},\Sigma_\alpha) \tag{3.47}$$

where $|\text{diag}(\cdot)|$ is the determinant of the diagonal matrix whose diagonal elements are $e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}$.

3.3.3.2 Enhancement of log Mel spectral vectors

To simplify the problem, Deng et. al. [55, 56, 57] assumed that the channel distortion is not significant and therefore can be ignored, i.e., $\mathbf{h} = 0$, and hence equation (3.47) is reduced to

$$p_y(\mathbf{y}|\mathbf{x},\mathbf{n}) = \frac{1}{2}|\text{diag}(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x})/2})|\mathcal{N}(\alpha(\mathbf{x},\mathbf{n},\mathbf{y}),\mathbf{0},\Sigma_\alpha) \tag{3.48}$$

It is not a Gaussian distribution, as the first term in equation (3.48) changes with $\mathbf{y}$.

To estimate $\mathbf{x}$, the noise $\mathbf{n}$ is first estimated by using a sequential estimation method,then the MMSE estimate of $\mathbf{x}$ can be found. It is noted that due to the non-Gaussian nature of $p_y(\mathbf{y}|\mathbf{x},\mathbf{n})$, the estimation process becomes very difficult. This problem is simplified by using the truncated second-order Taylor series to approximate the distribution (see [56] for detail).

The recognition performance of the phase-sensitive MMSE estimator [56] is found to be better than the that of the phase-insensitive MMSE estimator in [52] with 54% error rate reduction. This shows that the incorporation of the phase information benefits the feature compensation process by adding relevant information. If the phase factor is set to zero, the phase-sensitive MMSE estimator degenerates to spectral subtraction.

3.3.3.3 Incorporation of dynamic features

Later, the phase-sensitive MMSE estimator is expanded to include the first order derivatives of the speech features in the log Mel filterbank domain [55], due to the assumption that the strong dynamic property of speech features are important for the enhancement of the features. The static and dynamic features are assume to have GMM distribution and independent from each other. Then the noisy speech feature distribution function is derived and the clean speech features are estimated using the MMSE criterion.

The recognition accuracy on AURORA-2 shows that the incorporation of dynamic features leads to better performance. Furthermore, the enhanced spectrogram from system with use of dynamic features is smoother than that from system without use of dynamic features. The trend information in time introduced by dynamic features is orthogonal to the information of static features and therefore provides better enhancement.

3.3.3.4 Introducing the uncertainty of feature compensation

The work of Deng et. al. was further expanded by incorporating a feature compensation uncertainty [57] in the decoding process. The feature compensation uncertainty accounts for the deviation of the enhancement feature from the clean feature, i.e. the variance of the feature estimator. To better decode the noisy speech, this uncertainty should be taken into account in the decoding process. One way to do this is to integrate the acoustic score over this uncertainty space, i.e. over all possible clean feature values. One issue for incorporation of the uncertainty is how to efficiently calculate the integration. The integration is effectively the same as adding the variance of the feature estimator (the uncertainty) to the Gaussian's of the HMM states if the feature estimation error is assumed to be zero-mean Gaussian distribution [57]. Another issue is how to effectively estimate the feature estimator's variance. In [57], analytical solutions are derived by making use of the phase-sensitive environment model.

3.3.4 Missing Feature Approaches

One new group of robust feature compensation algorithms are based on the missing feature theory (MFT) [65, 66, 68], with the observation that the speech signal contains much redundant information, and human beings can recognize the words correctly with only partial speech information. This observation motivates a new way of recognizing the noisy speech—the recognition engine should only use features from the reliable spectral bins of the noisy speech signal, while the unreliable part should be discarded or processed before use. One way of classification of the features as reliable or unreliable is based on the local SNR of features with respect to time index and frequency bins due to the variability of the speech and noise. It is suggested that the features with high local SNR be labeled as reliable or present, while those with a low local SNR be labeled as

unreliable or missing. In the recognition process, the missing features are either discarded or re-estimated.

The MFT algorithms can be loosely classified into two groups: the first group requires recognition engine modifications and the second group does not. In [65], the algorithms requiring recognition engine modifications are called classifier compensation algorithms while the other with no engine modification are referred to as feature compensation algorithms.

Two classifier compensation algorithms that operate in the log Mel filterbank domain were proposed by Cooke et al [66]. The first algorithm, called the state-dependent imputation method, estimates the missing input features as the mean of the state-dependent distribution of the missing coefficients conditioned on the reliable coefficients during the recognition process. The second algorithm, called marginalization, performs the recognition solely based on the reliable coefficients. The performance of the imputation algorithm was improved in [67] by estimating the cepstral domain coefficients directly from the log Mel filterbank coefficients through a nonnegative least square approach.

Two feature compensation algorithms were also proposed by Raj [65]. Unlike the classifier compensation algorithms, the two feature compensation algorithms estimate the unreliable log Mel filterbank coefficients prior to MFCC generation. As this approach reconstruct all the filterbank coefficients prior to the recognition engine, no modification is required on the engine. The first algorithm, called the correlation-based reconstruction, estimates the filterbank coefficients of the missing components using the relationship between the missing components and the reliable neighbor components. The second algorithm, called the cluster-based reconstruction method, uses a Gaussian mixture model (GMM) to model the distribution of the log Mel filterbank coefficient vectors and reconstructs the missing components' filterbank coefficients using the estimated GMM. After the log Mel filterbank coefficients are reconstructed by either method, the MFCC are calculated and used as the feature for the recognition system. The major advantages of these two approaches are that no engine modification is required and that the recognition is performed in the cepstral domain.

Missing feature theory based approaches make no assumption about the noise, so they are robust for handling non-stationary noises. It is a promising technique and we will review the four existing techniques in chapter 4 in detail.

3.4 Model Adaptation Techniques

Unlike the feature compensation techniques, which aim at making the noisy feature as similar to the clean feature as possible, the model adaptation techniques, on the other hand, adapt the HMM acoustic model to the noisy acoustic environment. In this section, several model adaptation techniques are reviewed and compared. Furthermore, they are also compared with the feature compensation techniques.

3.4.1 PMC

Gales and Young [58] proposed the parallel model combination (PMC) approach, which compensates the acoustic model by combining it with a noise model in linear spectral domain. In their approach, the noise is represented by a single or multi-state HMM. As the noise and the speech are assumed to be independent and additive, the parameters of the acoustic model is compensated by adding the parameters of the noise model in linear spectral domain. Specifically, for each clean and noise state pair, the mean vectors and covariance matrices of the two model are combined using the following formulae

$$\hat{\mu} = g\mu + \tilde{\mu} \tag{3.49}$$

$$\hat{\Sigma} = g^2\Sigma + \tilde{\Sigma} \tag{3.50}$$

where $(\hat{\mu}, \hat{\Sigma})$, (μ, Σ) are the noisy and clean speech model parameters, $(\tilde{\mu}, \tilde{\Sigma})$ are the noise model parameters, all in linear spectral domain. The gain matching term g is used, as the relative strength of the speech and the noise in the testing environment may be different from that of the training environment, and it is estimated as

$$g = \frac{E_{ns} - E_n}{E_s} \tag{3.51}$$

where E_s, E_{ns} and E_n are the average energy of the clean speech, noisy speech and background noise respectively.

During the training process, the multi-state HMM for clean acoustic model and the single state HMM for noise model are trained in cepstral domain (MFCC). During the adaptation process, both the clean acoustic model and noise model are transform to linear spectral domain first, where they are combined to produce the noisy acoustic model, and

then transformed back to the cepstral domain again and used for recognition. Experiment results in [58] shows great improvement in recognition accuracy on isolated digit task. Furthermore, the using of multi-state HMM (2-4 states) for non-stationary produces much better result than that using single state HMM. However, PMC is computational expensive, as every model parameter needs to be adapted.

3.4.2 STAR

The STAR algorithm of Moreno [76, 51] is closely related to the RATZ algorithm (see section 3.3.2.2). Unlike the RATZ which uses a separate GMM for the prior distribution of clean speech, STAR utilizes the HMM in the CMU SPHINX-II speech recognition engine directly. The SPHINX-II speech recognition engine uses discrete HMM acoustic models and all the states share a pool of 256 Gaussians. Therefore, the acoustical model can also be seen as a GMM. STAR estimates the correcting terms, μ_k and Σ_k, for the GMM using the same way as RATZ, and then compensates the clean mean and variance to approximate the noisy speech distribution. As these Gaussians are shared by all HMM models, once these they are compensated, all the HMM states are adapted.

Although both the PMC and STAR add correction terms to the clean means and variances of the acoustic models, they achieve it in different ways. The PMC approach estimates the correction terms, i. e. the noise HMM, during the training process, while the STAR approach does this in the recognition process. The gain matching term g in PMC is computed using environment model, while the correction terms in STAR is data-driven and obtained by maximizing the likelihood of the observation in the Bayesian framework.

3.4.3 MLLR and MAP

Another two model adaptation methods, the maximum likelihood linear regression (MLLR) [59] and MAP [60, 61], are originally designed for adapting the speaker independent acoustical models to speakers. Due to the similarity between the speaker adaptation and environment adaptation, they are also used for noise robust speech recognition. In the following paragraphes, these two methods are discussed and compared.

The MAP approach adapts the acoustic model by optimally using the prior information in the clean trained HMM acoustical models and the posterior information in the noisy observations. The observations are recognized by the speech recognition, and only those with high acoustic likelihood score are used for adaptation. The Bayesian adaptation framework used in the MAP approach enables the optimal use of the noisy observations in model adaptation. When the adaptation data are few and the posterior information is weak compared to the prior HMM acoustical models, the models are not adapted much. As there are more and more adaptation data, the models becomes asymptotically equivalent to the ML estimate, which is provides optimal decision rule on the test data. However, this adaptation process is quite slow, as only the model parameters directly related to the adaptation data are adapted. In real applications, the adaptation data are few and hence it is necessary to reduce the number of model parameters needed to be adapted.

To achieve good adaptation performance, MLLR uses the parameter sharing strategy, i.e., the similar models are tied together and their parameters are adapted together. The degree of model tieing is high if the available amount of adaptation data is high and vice versa. For very few data, a global transform strategy maybe used. The basic MLLR adapts the mean vectors of the Gaussian by multiplying it with a transform matrix, which is obtained using maximum likelihood criterion and EM algorithm. The models tied together share the same transformation matrix. The advantage of the MLLR is its ability to provide good adaptation even if data are few. However, MLLR has poor asymptotic property, which leads to the fast saturation of performance gain with increased data. Usually, the MLLR outperforms the MAP if the adaptation data are few, but MAP adapts the models better when there are a lot of data. The combination of the two yields best performance.

3.5 Summary

In this chapter, we reviewed existing techniques for noise robust ASR in three groups, speech enhancement techniques, feature compensation techniques and the model adaptation techniques.

Speech enhancement techniques usually estimate the clean speech spectrum in the frequency domain. Spectral subtraction is an intuitive method and does not have optimality in statistical sense. The enhanced spectral vector may be not like real spectral vector at all, as it does not use prior speech information to constrain the estimation. Ephraim's estimator is the optimal estimator for the short time spectral amplitude of the clean speech in the MMSE sense, and it is also extended to operate in log Mel filterbank domain. Signal subspace methods are based on the observation that the clean speech signal does not span in all the dimensions of the speech.

The feature compensation techniques focus on removing the noise's effect from the noisy feature to minimize the mismatch between the training and testing process of speech recognition. The most simple way of removing the channel distortion is CMN, which subtracts the utterance mean vector from every frames of the utterance. CDCN extends CMN by using different correction vectors for different frames. The techniques from Deng et al. introduce the phase information in their environment model and therefore yields better performance. Unlike the CDCN and techniques of Deng et al. which use an environmental model, the RATZ method is a data-driven approach which learns the correction terms for the distribution parameters from the testing data directly.

The missing feature theory based techniques does not use explicitly defined environment model and has no assumption about the testing environment and noisy characteristics, so they can handle a wide range of signal corruptions. However, their performance is heavily dependent on correct estimation of data mask, which is a difficult problem.

The model based techniques modify the mean and variance of the Gaussians in the HMM to better represent the noisy speech. Some methods, such as PMC and STAR add a correction term to the mean and variance to compensate the noise and channel distortion effect. Another two speaker adaptation methods, the MAP method and MLLR method, are applied to environment adaptation with minor modifications due to the similarity between the two problems. MAP method is the principled way of adapting the acoustic model to the new observations, it has good asymptotic properties, but requires a relative larger amount of adaptation data for good adaptation. On the other hand, the MLLR method flexibly share a transformation matrix between similar acoustic models to reduce the number of parameters need to be estimated, which leads to smaller requirement for

the amount of adaptation data. However, it is not as good as MAP method when there are a large amount of adaptation data.

Chapter 4

Missing Feature Techniques

In this section, we will discuss the missing feature theory based noise robust ASR approaches in more detail. Specifically, the four basic approaches, namely the marginalization, state-dependent imputation, cluster-based reconstruction and correlation-based reconstruction will be examined and compared. We also proposed to cascade the cluster-based and correlation based reconstruction. The experimental results are shown and discussed.

4.1 Classifier Compensation Algorithms

The two MFT-based techniques from Cooke et al. [66] both require the modification to the recognition engine. The difference between the two algorithms is that the state-dependent imputation estimates the unreliable features before the calculation of the state likelihood score, while the marginalization totally ignore the unreliable features by integrating the likelihood over the space of the unreliable features. Before we go into the details of the algorithms, let's first establish the notations.

For simplicity, the log Mel filterbank spectral coefficient vector is called spectral vector in the following text. Let $\mathbf{y}(n)$ and $\mathbf{x}(n)$ denote the spectral vector of noisy speech and underlying clean speech of the n^{th} frame. The frame index n is usually dropped for a more compact mathematical representation. The elements of the noisy spectral vector $\mathbf{y}$ are rearranged to form two smaller vectors, namely the reliable component vector $\mathbf{y}_r$ and unreliable component vector $\mathbf{y}_u$, by comparing each component's local SNR to a specified threshold. The underlying clean spectral vector $\mathbf{x}$ for $\mathbf{y}$ is similarly decomposed

into two vectors, $\mathbf{x}_r$ and $\mathbf{x}_u$ whose respective elements corresponds to that of $\mathbf{y}_r$ and $\mathbf{y}_u$. It is assumed that the noisy reliable components $\mathbf{y}_r$ is a good approximation of the clean reliable components $\mathbf{x}_r$, while the noisy unreliable components $\mathbf{y}_u$ is an upper bound of the underlaying clean components $\mathbf{x}_u$ [65],

$$\mathbf{y}_r \approx \mathbf{x}_r \tag{4.1}$$

$$\mathbf{y}_u \geq \mathbf{x}_u \tag{4.2}$$

in component-wise sense. Equation (4.2) sets an upper bound on the estimated value of $\mathbf{x}_u$ and it is justified by the fact that the noisy components usually have a higher energy than the clean components if the noise and signal are uncorrelated.

4.1.1 State-Dependent Imputation

The state-dependent estimation of the unreliable features [66] $\mathbf{x}_u$ requires the probability distribution of the $\mathbf{x}_u$ conditioned on the i^{th} HMM state and reliable features $\mathbf{x}_r$:

$$f(\mathbf{x}_u|\mathbf{x}_r, C_i) = \frac{f(\mathbf{x}_u, \mathbf{x}_r|C_i)}{f(\mathbf{x}_r|C_i)} = \frac{f(\mathbf{x}|C_i)}{f(\mathbf{x}_r|C_i)} \tag{4.3}$$

where $f(\mathbf{x}_u, \mathbf{x}_r|C_i) = f(\mathbf{x}|C_i)$. The probability of $\mathbf{x}$ in a state is represented as

$$f(\mathbf{x}|C_i) = \sum_{k=1}^{M} P(k|C_i) f(\mathbf{x}|k, C_i) \tag{4.4}$$

where C_i denotes the i^{th} state, $P(k|C_i)$ is the prior weight of the k^{th} Gaussian in state C_i, M is the number of Gaussian mixtures and $f(\mathbf{x}|k, C_i) = \mathcal{N}(\mathbf{x}; \mu_i^k, \Sigma_i^k)$ represents the k^{th} Gaussian of the i^{th} state. Substitute equation (4.4) into (4.3), we get

$$f(\mathbf{x}_u|\mathbf{x}_r, C_i) = \frac{\sum_{k=1}^{M} P(k|C_i) f(\mathbf{x}_u, \mathbf{x}_r|k, C_i)}{f(\mathbf{x}_r|C_i)} \tag{4.5}$$

As the covariance matrix of each Gaussian are assumed to be diagonal, the reliable and unreliable features are independent in the mixture level, i.e. $f(\mathbf{x}_u, \mathbf{x}_r|k, C_i) =$

$f(\mathbf{x}_u|k, C_i)f(\mathbf{x}_r|k, C_i)$. Hence, equation (4.5) can be rewritten as

$$\begin{aligned} f(\mathbf{x}_u|\mathbf{x}_r, C_i) &= \frac{\sum_{k=1}^{M} P(k|C_i)f(\mathbf{x}_r|k, C_i)f(\mathbf{x}_u|k, C_i)}{f(\mathbf{x}_r|C_i)} \\ &= \frac{\sum_{k=1}^{M} f(\mathbf{x}_r, k|C_i)f(\mathbf{x}_u|k, C_i)}{f(\mathbf{x}_r|C_i)} \\ &= \sum_{k=1}^{M} \frac{f(\mathbf{x}_r, k|C_i)}{f(\mathbf{x}_r|C_i)} f(\mathbf{x}_u|k, C_i) \\ &= \sum_{k=1}^{M} P(k|\mathbf{x}_r, C_i)f(\mathbf{x}_u|k, C_i) \end{aligned} \tag{4.6}$$

This can be interpreted as a modified GMM for the unreliable features only, where the weight of the mixtures $P(k|\mathbf{x}_r, C_i)$ is the posterior weight conditioned on reliable features and state. Assume that the distribution $f(\mathbf{x}_u|\mathbf{x}_r, C_i)$ is unimodal, the imputation algorithm estimates the unreliable features $\mathbf{x}_u$ conditioned on the i^{th} state as its most likely value, i.e. the expected value [66]

$$\begin{aligned} \hat{\mathbf{x}}_{u,i} = E_{\mathbf{x}_u|\mathbf{x}_r,C_i}[\mathbf{x}_u|\mathbf{x}_r, C_i] &= \int \mathbf{x}_u f(\mathbf{x}_u|\mathbf{x}_r, C_i)d\mathbf{x}_u \\ &= \sum_{k=1}^{M} P(k|\mathbf{x}_r, C_i) \int \mathbf{x}_u f(\mathbf{x}_u|k, C_i)d\mathbf{x}_u \\ &= \sum_{k=1}^{M} P(k|\mathbf{x}_r, C_i)\mu_{\mathbf{x}_u|k,C_i} \end{aligned} \tag{4.7}$$

This estimate of $\mathbf{x}_u$ is just the weighted sum of the mean of $\mathbf{x}_u$ of the states distribution function. In run time, $\hat{\mathbf{x}}_{u,i}$ may become unrealistically large or small. In such situation, the original unreliable feature is adopted rather than the estimated one. In mathematical form, the $\hat{\mathbf{x}}_{u,i}$ is bound by

$$\tilde{\mathbf{x}}_{u,i} = \begin{cases} \hat{\mathbf{x}}_{u,i}, & \mathbf{x}_{low} \leq \hat{\mathbf{x}}_{u,i} \leq \mathbf{x}_{high}; \\ \mathbf{x}_{high}, & \hat{\mathbf{x}}_{u,i} > \mathbf{x}_{high}; \\ \mathbf{x}_{low}, & \mathbf{x}_{low} > \hat{\mathbf{x}}_{u,i}. \end{cases} \tag{4.8}$$

where $\mathbf{x}_{high}$ and $\mathbf{x}_{low}$ are our prior knowledge about the unreliable features. For log Mel spectral vectors, $\mathbf{x}_{high} = \mathbf{x}_u$ and $\mathbf{x}_{low} = 0$ is a reasonable choice.

During the decoding process, the unreliable features are estimated from the reliable features and the statistics of current state. Hence, there is one estimate for each state, and this is where the algorithm gets its name. Instead of the original unreliable features, the estimated features are used to calculate the state output probability. One advantage of of this approach is that the imputated features on the winning path, together with the reliable features, can be used for reconstruction of speech signal if needed.

4.1.2 Marginalization

The unreliable features $\mathbf{x}_u$ can be ignored by calculating the state output probability as

$$f(\mathbf{x}_r|C_i) = \int f(\mathbf{x}_u, \mathbf{x}_r|C_i)d\mathbf{x}_u \tag{4.9}$$

where the unreliable features are integrated off. Substitute equation (4.4) into (4.9), the following expression is obtained

$$\begin{aligned} f(\mathbf{x}_r|C_i) &= \sum_{k=1}^{M} P(k|C_i) f(\mathbf{x}_r|k, C_i) \int f(\mathbf{x}_u|k, C_i)d\mathbf{x}_u \\ &= \sum_{k=1}^{M} P(k|C_i) f(\mathbf{x}_r|k, C_i) \end{aligned} \tag{4.10}$$

where $\int f(\mathbf{x}_u|k, C_i)d\mathbf{x}_u = 1$. The above probability completely ignored the unreliable features, so the decision rule is not affected by unreliable features. However, if the bound information of the unreliable features are available, the classification can be improved by adding bound to the integration in equation (4.10). The bounded integration can be computed using the error function as follows

$$\int_{\mathbf{x}_{low}}^{\mathbf{x}_{high}} f(\mathbf{x}_u|k, C_i)d\mathbf{x}_u = \frac{1}{2}\left[\text{erf}\left(\frac{\mathbf{x}_{high,u} - \mu_{u|k,C_i}}{\sqrt{2}\sigma_{u|k,C_i}}\right) - \text{erf}\left(\frac{\mathbf{x}_{low,u} - \mu_{u|k,C_i}}{\sqrt{2}\sigma_{u|k,C_i}}\right)\right] \tag{4.11}$$

where erf is the error function and can be implemented using lookup table. The integration is integrated from $\mathbf{x}_{low,u}$ to $\mathbf{x}_{high,u}$ instead of $-\infty$ to ∞ to further constrain the probability.

According to the results presented in [66], marginalization generally performs better than the state-based imputation. However, state-dependent imputation produces the

reconstructed spectrogram, which can be used for further processing, such as calculating the MFCC via DCT. It is well known that ASR system based on MFCC performs much better than that based on log Mel filterbank coefficients. One problem for the imputation method is the assumption of unimodal for the distribution $f(\mathbf{x}_u|\mathbf{x}_u, C_i)$. If it is multimodal, the best estimate of unreliable features need to be chosen form the modes. Without other information, it may be reasonable to choose the mode with the highest probability. Another problem is that even if $f(\mathbf{x}_u|\mathbf{x}_u, C_i)$ is unimodal, the imputation scheme discussed above is an ad-hoc approach and no optimality in any form can be claimed.

4.1.3 Extension to Operate in the Cepstral Domain

Hugo [67, 68] extends the above approaches to operate on the cepstral domain. The $2M+1$ frames of K dimension log Mel spectral vectors centered at frame n are concatenated together to form a vector $\mathbf{x}'(n)$:

$$\mathbf{x}'(n) = \begin{bmatrix} \mathbf{x}(n-M) \\ \vdots \\ \mathbf{x}(n) \\ \vdots \\ \mathbf{x}(n+M) \end{bmatrix} \tag{4.12}$$

where the spectral vectors are assumed to be column vectors. The cepstral feature vector $\mathbf{s}(n)$ is obtained by some transformation:

$$\mathbf{s}(n) = C\mathbf{x}'(n) \tag{4.13}$$

where the transformation matrix C's dimension is $D \times (2M+1)K$, and D is the dimension of the feature vector $\mathbf{s}$. C is defined as

$$C = \begin{bmatrix} \mathbf{b}_{static} \\ \mathbf{b}_{delta} \\ \mathbf{b}_{acc} \end{bmatrix} \otimes C_{DCT} \tag{4.14}$$

where C_{DCT} is the $(D/3) \times K$ truncated DCT matrix and $\otimes$ denotes the Kronecker product (see Appendix A.1 for details of Kronecker product). The $2M+1$ dimension row vectors $\mathbf{b}_{static}$, $\mathbf{b}_{delta}$ and $\mathbf{b}_{acc}$ are the FIR filter coefficients to compute the static, delta

and acceleration features. For a specific Gaussian mixture in a specific state, the $\mathbf{x}(n)$ can be estimated by maximizing the log likelihood of $\mathbf{s}(n)$ subjected by the constraints in equation (4.2):

$$\begin{aligned} \mathbf{x} &= \arg\max_{\mathbf{x}_u} \log f(\mathbf{s}|i,k) \\ &= \arg\min_{\mathbf{x}_u} (C\mathbf{x}-\mu_{ik})^T \Sigma_{ik}^{-1} (C\mathbf{x}-\mu_{ik}) \end{aligned} \tag{4.15}$$

subject to

$$\mathbf{x}_u \leq \mathbf{y}_u \text{ and } \mathbf{x}_r = \mathbf{y}_r \tag{4.16}$$

where $f(\mathbf{s}|i,k)$ is the Gaussian distribution of $\mathbf{s}$ of current mixture, i and k represents the state index and mixture index respectively. This problem is further converted to a non-negative least square problem and solved in [67].

4.2 Feature Compensation Algorithms

This section briefly introduces the two feature compensation algorithms of [65], specifically the correlation based and clustered based reconstruction approaches.

4.2.1 Correlation-Based Reconstruction

In the correlation based reconstruction method, the spectral vectors of the noisy utterances are assumed to be from a single stationary multivariate Gaussian random process. By estimating the cross-covariance of the spectral components for various lags, we can capture the relationship between the current frame's components and its neighbors. Given the trained covariance matrices, the reconstruction of the noisy vectors with missing features can be effected. As described in [65], the cross-covariance of any two components in the spectrogram is captured as $c(n_1, n_2, k_1, k_2)$ where n_1, n_2 denote frame index of the two components being compared and k_1, k_2 denote the Mel frequency bin of the respective components. Assuming the source is stationary, we have

$$c(n_1, n_2, k_1, k_2) = c(n_1 - n_2, k_1, k_2) = c(\Delta n, k_1, k_2) \tag{4.17}$$

and the relative cross-covariance, or correlation coefficient, between two components can be estimated as follows.

$$r(\Delta n, k_1, k_2) = \frac{c(\Delta n, k_1, k_2)}{\sqrt{c(0, k_1, k_1)c(0, k_2, k_2)}} \tag{4.18}$$

The mean and covariance are estimated from training data. During the reconstruction phase, the unreliable spectral components is estimated from the reliable components in both the current and neighboring frames. The unreliable spectral components of the current frame are organized to form vector $\mathbf{y}_u$ and all the reliable components in the spectrogram which are related to any component of $\mathbf{y}_u$ are organized to form the vector $\mathbf{y}'_r$. Let $\mathbf{x}'_r$ and $\mathbf{x}_u$ be the underlying true value for $\mathbf{y}'_r$ and $\mathbf{y}_u$ respectively. The criterion to judge whether two spectral components are related is based on whether their relative cross-covariance is larger than a preset threshold. Because of the Gaussian assumption for $\mathbf{x}$, the joint distribution $f(\mathbf{x}'_r, \mathbf{x}_u)$ of $\mathbf{x}'_r$ and $\mathbf{x}_u$ is also a Gaussian distribution whose mean and covariance matrix can be generated from the statistics of $\mathbf{x}'_r$ and $\mathbf{x}_u$ [37]. After the $f(\mathbf{x}'_r, \mathbf{x}_u)$ is obtained, the conditional probability distribution of the unreliable features $f(\mathbf{x}_u|\mathbf{x}'_r)$ can be found by

$$f(\mathbf{x}_u|\mathbf{x}'_r) = \frac{f(\mathbf{x}_u, \mathbf{x}'_r)}{f(\mathbf{x}'_r)} \tag{4.19}$$

and the $\mathbf{x}_u$ can be found by

$$\hat{\mathbf{x}}_u = \arg\max_{\mathbf{x}_u} f(\mathbf{x}_u|\mathbf{x}'_r) \tag{4.20}$$

By using the relationship in equation (4.1,4.2), the estimates of the unreliable components are obtained by maximizing the following function

$$\hat{\mathbf{x}}_u = \arg\max_{\mathbf{x}_u} f(\mathbf{x}_u|\mathbf{x}_u \leq \mathbf{y}_u, \mathbf{x}'_r = \mathbf{y}'_r) \tag{4.21}$$

Using Bayes rule and denoting $\mathbf{x}'_r = \mathbf{y}'_r$ as $\mathbf{y}'_r$, the above equation can be rewritten as

$$\hat{\mathbf{x}}_u = \arg\max_{\mathbf{x}_u} f(\mathbf{x}_u, \mathbf{x}_u \leq \mathbf{y}_u|\mathbf{y}'_r) \tag{4.22}$$

An iterative approach was proposed in [65] to solve the above Maximum A Posteriori (MAP) problem.

4.2.2 Cluster-Based Reconstruction

The cluster-based reconstruction assumes that the spectral vectors of the clean utterances are from an independent, identically distributed multivariate random process (IID) which

can be modeled by a GMM [65]. Therefore, the probability distribution of a spectral vector $\mathbf{x}$ is

$$f(\mathbf{x}) = \sum_{k=1}^{K} P(k) f(\mathbf{x}|k) \tag{4.23}$$

where $P(k)$ is the *a priori* probability of the k^{th} cluster and the distribution of the k^{th} cluster $f(\mathbf{x}|k)$ is defined as

$$P(\mathbf{x}|k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp(-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x} - \mu_k)) \tag{4.24}$$

where d is the dimension of the spectral vector and μ_k and Σ_k are the mean vector and covariance matrix of the k^{th} cluster respectively. Note that the IID assumption inherently ignores any information between the neighboring vectors.

In the training phase, the cluster parameters, including the mean vector, covariance matrix and *a priori* probability are estimated using Expectation-Maximization (EM) method or K-mean clustering algorithm. In the reconstruction phase, the *a posteriori* probability of the cluster given the noisy vector is calculated by

$$P(k|\mathbf{y}_r, \mathbf{x}_u \leq \mathbf{y}_u) = \frac{P(k) f(\mathbf{y}_r, \mathbf{x}_u \leq \mathbf{y}_u|k)}{\sum_{j=1}^{K} P(j) f(\mathbf{y}_r, \mathbf{x}_u \leq \mathbf{y}_u|j)} \tag{4.25}$$

and

$$f(\mathbf{y}_r, \mathbf{x}_u \leq \mathbf{y}_u|k) = \int_{-\infty}^{\mathbf{y}_u} f(\mathbf{y}_r, \mathbf{x}_u|k) dX_u \tag{4.26}$$

Note that the $f(\mathbf{y}_r, \mathbf{x}_u \leq \mathbf{y}_u|k)$ uses the relationship in equation (4.1,4.2), and that $P(\mathbf{y}_r, \mathbf{x}_u|k) = P(\mathbf{x}|k)$. For computational simplicity, the covariance matrix of $P(\mathbf{y}_r, \mathbf{x}_u|k)$ is assumed to be diagonal when calculating equation (4.26). This is done by ignoring the off-diagonal elements of the covariance matrix of $P(\mathbf{x}|k)$.

After obtaining the posterior weight of the clusters $P(k|\mathbf{y}_r, \mathbf{x}_u \leq \mathbf{y}_u)$, the unreliable components are estimated using every cluster, and the final estimate is the weighted sum of all the cluster-dependent estimates. The estimate of the noisy vector for the k^{th} cluster is calculated as

$$\hat{\mathbf{x}}_u^k = \arg\max_{\mathbf{x}_u} f(\mathbf{x}_u|k, \mathbf{x}_u \leq \mathbf{y}_u, \mathbf{x}_r = \mathbf{y}_r) \tag{4.27}$$

Similar to the correlation based reconstruction, this is equivalent to

$$\hat{\mathbf{x}}_u^k = \arg\max_{\mathbf{x}_u} f(\mathbf{x}_u, \mathbf{x}_u \leq \mathbf{y}_u|k, \mathbf{y}_r) \tag{4.28}$$

This is essentially the same MAP problem as that in the correlation based reconstruction. The equation (4.28) can be solved using the same iterative MAP procedure as the one used in correlation based reconstruction method. The final estimate of the unreliable spectral vector is obtained by

$$\hat{\mathbf{x}}_u = \sum_{k=1}^{K} P(k|\mathbf{y}_r, \mathbf{x}_u \leq \mathbf{y}_u)\hat{\mathbf{x}}_u^k \tag{4.29}$$

Comparing correlation based reconstruction and cluster based reconstruction, we found that the major difference between the two is the source of information used, not the method of estimation. The correlation based reconstruction utilizes the inter-frame statistics while the cluster based reconstruction uses the intra-frame statistics. The MAP estimation procedure is the same for the two methods, once the joint probability of the reliable and unreliable features are found.

From the results reported by Raj [65, 64], the cluster-based reconstruction is superior to the correlation based reconstruction, except when the locations of the unreliable features are evenly distributed in the spectrogram. This may be due to the fact that the non-stationary characteristics of spectral vectors does not affect the GMM modeling in the cluster based reconstruction, but it conflicts with the assumption of stationarity in the correlation based approach. The single model for correlation-based method does not model the dynamics of the speech enough, as there are many different phonetic contexts in speech.

4.2.3 Comparison to Classifier Compensation Algorithms

The advantage the classifier compensation approaches is that they enjoy the statistics of the speech recognition to either estimate the missing features or marginalize them. Feature compensation approaches which uses either Gaussian or GMM to model the clean speech statistics is believed to be less effective than classifier compensation approach. However, the use of statistics in the speech recognition requires modifications to decoding algorithm in the engine, which is a weakness. On the other hand the feature compensation approaches reproduce the whole spectrogram prior to recognition, so it can be applied to any standard recognition engine. Another disadvantage of classifier compensation approaches is that the log Mel filterbank feature is used, which is not as good

as the cepstral feature used in the feature compensation approaches. Although Hugo [68] expanded the classifier compensation approaches into cepstral domain, the proposed method is computational complex.

4.3 Data Mask Estimation

The data mask that identifies the reliable and unreliable speech features is crucial for the overall performance of the missing feature based techniques. There are three groups of techniques for obtaining this mask. The first group of techniques depends on the local SNR of the speech feature. A high local SNR indicates the cleanness and high reliability of the feature and vice versa. Another technique treats the masking as a two-class classification problem, where the speech features are classified into either the reliable class or the unreliable class, using some derived features. In the third group of techniques, the perceptual criteria are used to find the mask, usually through computational acoustic scene analysis (CASA).

4.3.1 Local SNR-Based Methods

The local SNR approaches marks a feature as unreliable or missing, if its local SNR is less than a predefined threshold

$$\mathrm{SNR}_{local} = 20 \log \frac{\hat{x}}{\hat{n}} < \mathrm{Threshold} \tag{4.30}$$

where $\hat{x}$ and $\hat{n}$ is the estimate of the clean speech energy and the noise energy in current spectrogram location. It is obvious that this method largely depends on the noise estimation.

When the true clean speech energy and noise energy are used in the evaluation of the local SNR, the mask obtained is called the oracle mask or *a priori* mask. This is only achievable when we know the uncorrupted clean speech signal and the corrupting additive noise signal prior to the mixing of the two. Therefore, it is usually used as a benchmark to identify the potential of the missing future algorithm. Although oracle mask allows us to evaluate the missing feature algorithm performance, it may not be the best mask. Cooke [66] suggested that better mask may be achieved by using a variable

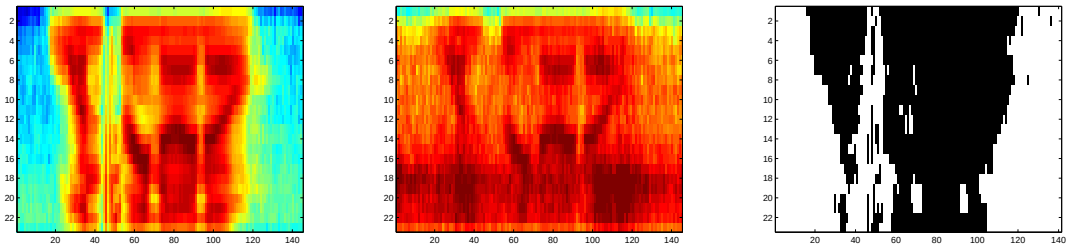

Figure 4.1: An illustration of oracle data mask in the log Mel filterbank domain. Left: Clean speech for utterance "one-three-nine-oh"; Center: Noisy speech corrupted by subway noise with SNR=5dB; Right: Oracle data mask with SNR threshold = -5dB—white cells represent unreliable features and black cells represent reliable features

local SNR threshold dependent on the amount of missing data. The figure 4.1 shows a example of oracle mask obtained using the local SNR criterion.

A closed related mask generating algorithm is the negative energy criterion, which is motivated by spectral subtraction. Due to the variance of noise estimation, spectral subtraction produces negative energy in some spectrogram locations, which is called "musical noise". In these locations, the noise is assumed to be dorminant, so they are marked as unreliable.

$$|y| - |\hat{n}| < 0 \tag{4.31}$$

where $|y|$ and $\hat{n}$ are the energy of the noisy speech and the noise in the current spectrogram location.

In [66], best mask estimate are reported by using both the local SNR criterion (equation (4.30)) and the negative energy criterion (equation (4.31)). Generally, the local SNR based methods works well if the corrupting noise is stationary or quasi-stationary. However, when the noise is non-stationary, the mask obtained using this criterion can be very inaccurate.

4.3.2 Bayesian Classifier Approach

Instead of using the local SNR as the only information to judge whether a speech feature is reliable or not, Seltzer [69] proposed to incorporate more information for the mask generation. In Seltzer's approach, the mask generation is treated as a statistical classification problem using Bayesian's decision rule, where the features are derived for each

spectrogram location and the two output classes are reliable and unreliable. The features for each spectrogram location include: a) the log ratio of the energies of the speech signal in the peak and valley of the formant, b) the ratio of the first and second autocorrelation peak of the signal within the frame, c) the ratio of sub-band energy to full-band energy, d) the ratio of sub-band energy to full-band noise floor, e) the ratio of sub-band energy to sub-band noise floor and f) the flatness of the spectral vector. As the unvoiced frames does not have formant and pitch, feature a and b are not used for them.

The training of the classifier is described as follows: The training data are first segmented into voiced and unvoiced segments, and then the features for every spectrogram location are computed. After that, the oracle data mask obtained using local SNR criterion is used to train the classifiers for voiced and unvoiced speech separately. The distribution of the feature vectors for all frequency bands are modeled as GMM. For every frequency band, two classifiers are trained, one for voiced speech and the other for unvoiced speech. The parameters of the Guassian mixtures that represents the distribution of the feature vectors of the reliable locations in the k^{th} frequency band are estimated from all the reliable voiced samples in the k^{th} frequency band in the training data. This is also true for other frequency band, unreliable location and unvoiced speech.

In the mask generating process, the noisy speech is first separated into voiced and unvoiced segments. The feature vector for each location is computed and the classification is based on the following decision rule

$$(m,k) \text{ is } \begin{cases} \text{reliable}, & \text{if } P_v^k(\text{r})P_v^k(F(m,k)|\text{r}) > P_v^k(\text{u})P_v^k(F(m,k)|\text{u}); \\ \text{unreliable}, & \text{otherwise.} \end{cases} \tag{4.32}$$

where (m,k) represents location in the m^{th} frame and k^{th} frequency band, v is the tag specifying voiced and unvoiced frames, $P_v^k(\text{r})$ and $P_v^k(\text{u})$ are the prior probability of the specified location to be reliable and unreliable respectively, and $P_v^k(F(m,k)|\text{r})$ and $P_v^k(F(m,k)|\text{u})$ represents the distribution of the feature vectors of the reliable and unreliable locations respectively.

The Bayesian classifier based approach is superior to the local SNR based approaches, as it utilizes other information besides the local SNR. It is more robust in the non-stationary noise situation, as it does not heavily depends on accurate estimate of the noise. In [69], it is reported that the mask from Bayesian classifier approach is effective

for music corrupted speech, while the mask from SNR criterion totally fails. On the other hand, as the Bayesian classifier approach utilizes many speech specific features, its performance is expected to be bad when the corrupting noise has similar characteristics as the speech.

4.3.3 Perceptual Criteria-Based Masks

It is suggested that the structure information in speech signal may be useful in identifying the missing features. For example, in [71], all the speech features at the harmonics of the pitch may be assumed to be reliable if the current frame is voiced and has a valid pitch estimate. This strategy is better used together with other mask identification techniques, such as the one based on local SNR estimate. In another example, Palomaki et. al. [72] proposed to use a binaural processing model to extract the information about the inter-aural time delays and intensity differences for identifying the reliable time-frequency regions. This approach leads to significant improvement in speech recognition accuracy when the target speech signal and the noise signal come from different azimuths.

As the perceptual criteria-based masks exploit the inherent structure information of the speech signal itself, they are more robust than the previous discussed methods. They are effective in different types of noise, even when the noise is a competing speech signal.

4.3.4 Soft Decision Mask

Instead of using hard decision data mask, soft decision masks, where the reliability of the spectrogram locations are represented using continuous value, are proposed. Barker et. al. [70] proposed to estimate the reliability γ by

$$\gamma \approx \frac{1}{1 + \exp(-\alpha(\mathrm{SNR}(m,k) - \beta)} \tag{4.33}$$

where α and β are typically 3.0 and 0.0 respectively. For the Bayesian classifier approach, γ can be obtained from the posterior probabilities of the location to be reliable and unreliable.

Although soft decision mask is conveys more information than the hard decision mask, only the bound marginalization method can effectively take advantage of it with minor modifications. Other missing feature based techniques require unambiguous identification of unreliable spectral components.

Chapter 5

Vector Autoregressive Modeling of Speech Feature Vectors

The speech signal is a slowly time-varying signal. The slow time-varying nature is reflected by highly correlated spectral frames, in other words, speech frames are highly predictive. However, current modeling of speech spectral vectors, such as hidden Markov model (HMM) and Gaussian mixture model (GMM) [64][65] usually assume that the spectral vectors of neighboring frames are independent in order to achieve mathematical simplicity. The HMM framework captures the relationship between neighboring frames weakly using the transition probabilities, while GMM completely discards this relationship. It is believed that if the inter-frame statistics are captured and harnessed properly, the performance of both spectral vector reconstruction and speech recognition will be improved. In this chapter, we propose a new way of modeling the speech spectral vectors, that is the vector autoregressive model. Two feature compensation frameworks are proposed based on the VAR model, and experiments were carried out to examine their performance. Results show that the proposed frameworks are more effective than the cluster based reconstruction discussed in previous chapter.

5.1 Apply VAR Model to Speech Feature Vectors

5.1.1 Vector Autoregressive Model

Vector autoregressive is a special model for multiple time series analysis. To explain the VAR, let's begin with time series analysis. In many applications, we need to forecast

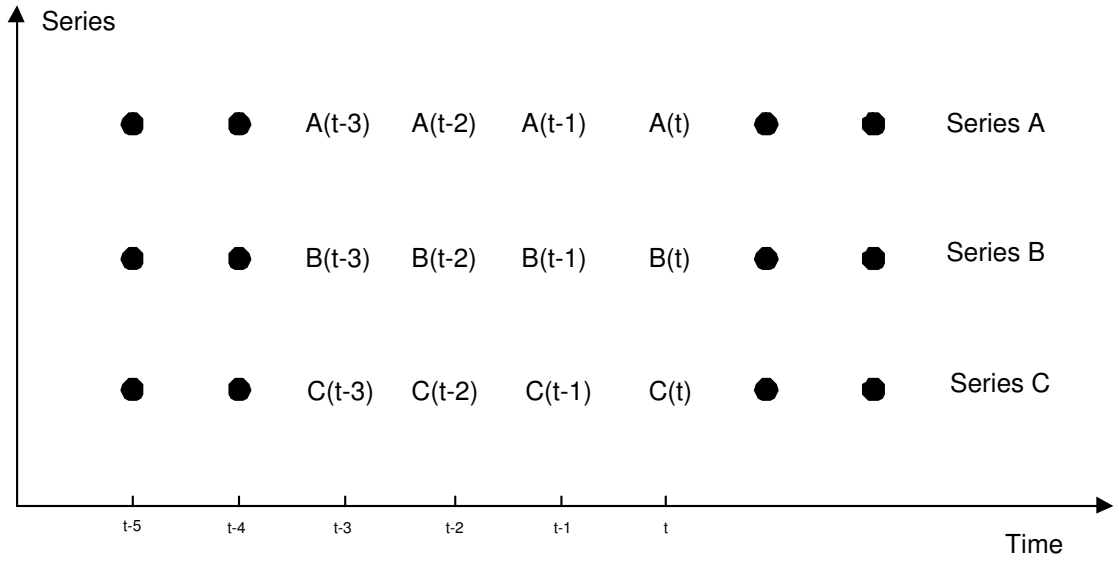

Figure 5.1: An illustration of multiple time series prediction.

some values in the future, such as the weather forecast or economic growth prediction. It is good to predict these values from their past values. This is because there may be some relationship between the past value and the future value of these processes. For example, the economic growth rate of one country in next year is probably quite close to the growth rate of current year. Sometimes, the value to be predicted is not only related to its past values, but also related to the past values of other processes, then we get multiple time series analysis.

Suppose we have three related time series, as shown in Fig.5.1. The value of $A(t)$ can be predicted from the past value of all A, B and C processes:

$$A(t) = f(A(t-1), ..., A(t-P); B(t-1), ..., B(t-P); C(t-1), ..., C(t-P)) \quad (5.1)$$

where P decides how many past values are considered in the prediction. The simplest format of prediction is the linear prediction:

$$\begin{aligned} A(t) &= \alpha^A_{t-1}A(t-1) + ... + \alpha^A_{t-P}A(t-P) + \beta^A_{t-1}B(t-1) + ... + \beta^A_{t-P}B(t-P) \\ &\quad +\gamma^A_{t-1}C(t-1) + ... + \gamma^A_{t-P}C(t-P) + e^A_t \end{aligned} \quad (5.2)$$

where e^A_t is the prediction error and is white noise, α^A_{t-1} denotes the weight of $A(t-1)$ for prediction of $A(t)$, and similar for β and γ. To put the equation (5.2) into matrix

and vector format, let's first define $\mathbf{x}_t = [A(t), B(t), C(t)]^T$ be the vector of observation, where $[\cdot]$ denotes the vector concatenation, and $\mathbf{w}_{A,t} = [\alpha_t^A, \beta_t^A, \gamma_t^A]^T$ be the vector of weights for prediction of $A(t)$. The linear prediction of $A(t)$ can be rewritten as

$$A(t) = \mathbf{w}_{A,t-1}^T \mathbf{x}_{t-1} + ... + \mathbf{w}_{A,t-P}^T \mathbf{x}_{t-P} + e_t^A \tag{5.3}$$

Furthermore, the whole observation vector of time t can be predicted as

$$\mathbf{x}(t) = \mathrm{W}_{t-1}^T \mathbf{x}_{t-1} + ... + \mathrm{W}_{t-P}^T \mathbf{x}_{t-P} + \mathbf{e}_t \tag{5.4}$$

where the weight matrix is obtained by concatenate the weight vectors for prediction of $A(t)$, $B(t)$ and $C(t)$, $\mathrm{W}_t = [\mathbf{w}_{A,t}, \mathbf{w}_{B,t}, \mathbf{w}_{C,t}]$, and the error vector $\mathbf{e}_t = [e_t^A, e_t^B, e_t^C]^T$ is the concatenation of errors from prediction of $A(t)$, $B(t)$ and $C(t)$.

5.1.2 Modeling Speech Feature Vectors

This section describes the mathematical representation of the log Mel spectral vectors of speech in the VAR model. We use the VAR models to capture the inter-frame relationship between the spectral vectors such that the current frame feature can be predicted by its past (forward prediction) or future vectors (backward prediction) as illustrated in Fig. 5.2. The elements of the log Mel spectral vector $\mathbf{x}(n)$ are predicted as a linear combination of all the elements of either the past vectors or the future vectors plus a prediction error that is white noise. The mathematical derivation of forward prediction and backward prediction are similar, hence only the derivation of the forward prediction model is discussed. The j^{th} element of $\mathbf{x}(n)$ can be predicted as [15].

$$x_j(n) = -\sum_{i=1}^{P} \mathbf{w}_{i,j}^T \mathbf{x}(n-i) + e_j(n) \tag{5.5}$$

where $\mathbf{w}_{i,j}$, for $i = 1, ..., P$ are the $D \times 1$ weight vectors, D is the dimension of $\mathbf{x}(n)$, $e_j(n)$ is a white noise or innovation process. The minus sign is there to follow the convention in linear prediction literature. Let $\mathrm{W}_i = [\mathbf{w}_{i,1}, ..., \mathbf{w}_{i,D}]$, $\mathbf{x}(n) = [x_1(n), ..., x_D(n)]^T$ and $\mathbf{e}(n) = [e_1(n), ..., e_D(n)]^T$, Eq (5.5) can be rewritten as:

$$\mathbf{x}(n) = -\sum_{i=1}^{P} \mathrm{W}_i^T \mathbf{x}(n-i) + \mathbf{e}(n) \tag{5.6}$$

where W_i is the $D \times D$ weight matrix for the i^{th} order, $\mathbf{e}(n)$ is the D dimensional white noise.

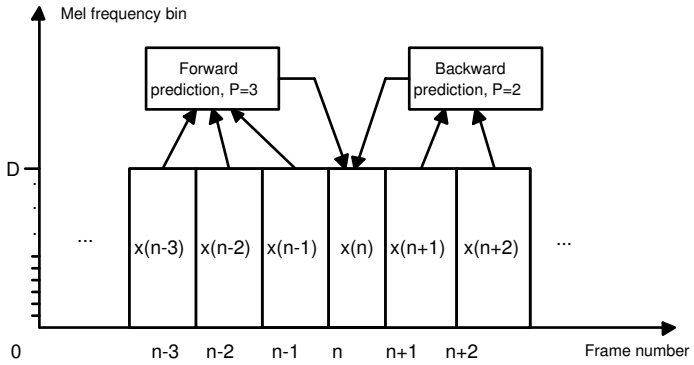

Figure 5.2: The illustration for forward and backward prediction of speech feature vectors.

5.1.3 Estimating the Model Parameters

Here we describe how to find the Least Square solution of W_i [15]. Let $\mathbf{x}(n-i)$ for $i = 1, ..., P$ be concatenated to form a super vector $\mathbf{o}(n)$, i.e. $\mathbf{o}(n) = [\mathbf{x}(n-1)^T, ..., \mathbf{x}(n-P)^T]^T$, and let W_i be concatenated to form $\mathrm{B} = [-\mathrm{W}_1^T, ..., -\mathrm{W}_P^T]$. Then Eq (5.6) can be rewritten as

$$\mathbf{x}(n) = \mathrm{B}\mathbf{o}(n) + \mathbf{e}(n) \tag{5.7}$$

To estimate the weight matrix B, we form training samples $\mathbf{r}(n)$ as the concatenation of the desired vector $\mathbf{x}(n)$ and the super input vector $\mathbf{o}(n)$, i.e., $\mathbf{r}(n) = [\mathbf{x}(n)^T, \mathbf{o}(n)^T]^T$. Suppose we have a collection of M training samples, denote the input vector $\mathbf{o}(n)$ of all the samples as $\mathrm{O} = [\mathbf{o}(1), ..., \mathbf{o}(M)]$, and the corresponding desired vectors $\mathbf{x}(n)$ as $\mathrm{X} = [\mathbf{x}(1), ..., \mathbf{x}(M)]$. The Least Square solution can be found by the following equation:

$$\hat{\mathrm{B}} = \mathrm{X}\mathrm{O}^T(\mathrm{O}\mathrm{O}^T)^{-1} = \mathrm{X}\mathrm{O}^+ \tag{5.8}$$

where $\hat{\mathrm{B}}$ is the estimate of B and O^+ is the pseudoinverse of O. The weight matrix $\hat{\mathrm{B}}$ is used to predict the spectral vectors during reconstruction phase, using the formula

$$\hat{\mathbf{x}}(n) = \hat{\mathrm{B}}\mathbf{o}(n) \tag{5.9}$$

where $\hat{\mathbf{x}}(n)$ is the estimate of $\mathbf{x}(n)$.

5.1.4 Modeling the Non-Stationarity of Speech

Given a collection of training samples, we have described how to construct a VAR model. However, we know that speech signal is not a stationary process. We can not expect to

use one VAR model for spectral feature prediction in many different phonetic contexts. Note that speech is composed of a finite number of phonemes. Studies have shown that speech signals of the same phoneme share similar spectral pattern. One solution to vector autoregressive modeling for a ASR task is to have multiple VAR models according to the short-term spectral characteristics. Each VAR is modeled by a collection of homogeneous spectral training samples.

5.2 Feature Compensation Schemes

This section describes the detailed implementations of the proposed VAR based missing feature reconstruction schemes. We propose two schemes, one uses clean training data, and the other uses noisy training data corrupted by white noise. We present details only for the forward prediction model as the procedure for backward prediction model are similar except the direction of prediction.

5.2.1 Scheme I: Use Clean Trained VAR Model

Under missing feature framework, the unreliable features are discarded, and their values are estimated. In our VAR-based approach, the values of these unreliable features are predicted from their neighboring frames. In the training process, the VAR models are estimated, and in the testing process, the unreliable features are predicted according to these VAR models. The training approach is to train the models on clean speech data. The training and testing process are summarized in next two sections.

5.2.1.1 Training procedures

The objective of the training process is to cluster the training samples and estimate the parameters of the VAR models for every cluster. The procedure for clean VAR model training is summarized as follows (see Fig. 5.3.A):

(i) Collect all the training samples $\mathbf{r}(n) = [\mathbf{x}(n)^T, \mathbf{o}(n)^T]^T$ where n denotes the n^{th} frame in the the training set.

(ii) Use K-means algorithm to cluster all $\mathbf{r}(n)$ into K clusters, called spectral clusters. Estimate the mean vector and covariance matrix of the Gaussian models $\mathcal{N}(\mathbf{r}, \mu_k, \Sigma_k)$ for each cluster $k = 1, ..., K$.

(iii) Label each sample with a cluster id $k = 1, ..., K$. For each cluster, collect all the input vectors $\mathbf{o}(n)$ and their corresponding desired vectors $\mathbf{x}(n)$, then estimate the weight matrix B_k of cluster k using Eq (5.8).

5.2.1.2 Testing procedures

The procedure of estimating missing features is as follows (Fig. 5.3B):

(i) For each noisy vector $\mathbf{x}(n)$, identify the set of reliable and unreliable features. We use the oracle mask to do so in this paper. For each of the unreliable features, go through step 2-4;

(ii) Estimate the spectral vector $\mathbf{x}(n)$ using the VAR models, we first form a vector $\mathbf{r}(n) = [\mathbf{x}(n)^T, \hat{\mathbf{x}}(n-1)^T, ..., \hat{\mathbf{x}}(n-P)^T]^T$, where the first vector is the noisy vector $\mathbf{x}(n)$ and
$\{\hat{\mathbf{x}}(n-j)\}_{j=1}^{P}$ are the reconstructed past vectors.

(iii) Find the *a posteriori* probability for every Gaussian model $\mathcal{N}(\mathbf{r}, \mu_k, \Sigma_k)$ for $k = 1, ..., K$ given vector $\mathbf{r}(n)$ in the same manner as in the cluster-based method reported in [65].

$$p(\mathbf{r}(n); k) = \mathcal{N}(\mathbf{r}(n), \mu_k, \Sigma_k), \quad k = 1, ..., K \tag{5.10}$$

$$p(k|\mathbf{r}(n)) = p(\mathbf{r}(n); k) / \sum_{l=1}^{C} p(\mathbf{r}(n); l) \tag{5.11}$$

where $p(\mathbf{r}(n); k)$ is the likelihood of $\mathbf{r}(n)$ on the k^{th} Gaussian model and $p(k|\mathbf{r}(n))$ is the *a posteriori* probability.

(iv) Form super input vector $\mathbf{o}(n) = [\hat{\mathbf{x}}(n-1)^T, ..., \hat{\mathbf{x}}(n-P)^T]^T$. Find the model-dependent prediction $\tilde{\mathbf{x}}_k(n)$ of $\mathbf{x}(n)$ using

$$\tilde{\mathbf{x}}_k(n) = \mathrm{B}_k \mathbf{o}(n), \quad k = 1, ..., K \tag{5.12}$$

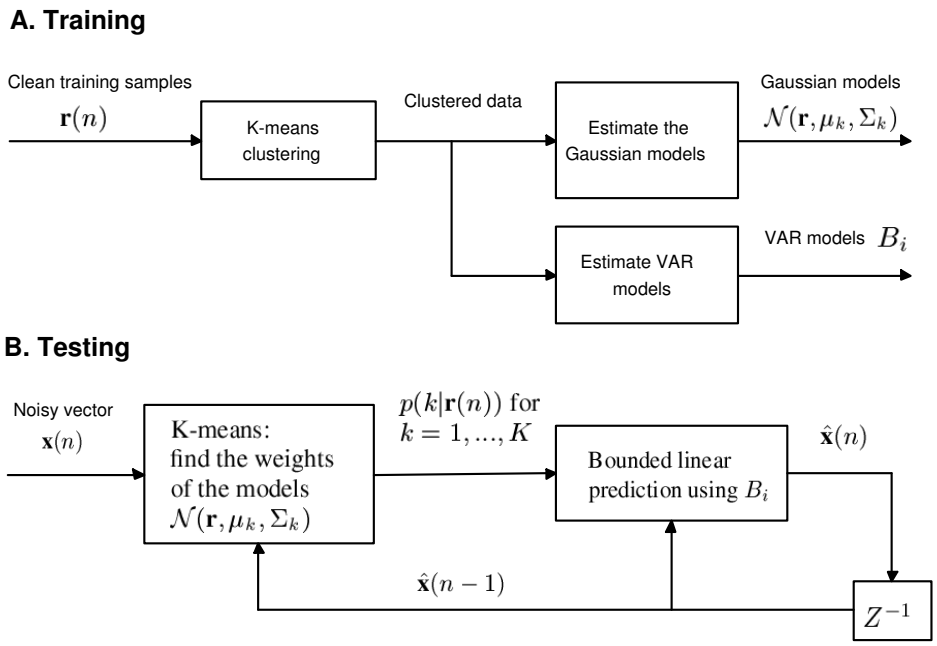

Figure 5.3: Feature compensation scheme I: Use the clean-trained VAR models

(v) Reconstruct the feature vector $\mathbf{x}(n)$ by

$$\hat{x}_j(n) = \begin{cases} min\{x_j(n)\ , \sum_{k=1}^{K} p(k|\mathbf{r}(n))\tilde{x}_{kj}(n)\ \}, \\ \quad \text{for unreliable features;} \\ x_j(n), \text{ for reliable features.} \end{cases} \tag{5.13}$$

where the reliable features keeps original values while the unreliable features are replaced by the weighted sum bounded to the corresponding noisy values, and $\tilde{x}_{kj}(n)$ is the j^{th} element of $\tilde{\mathbf{x}}_k(n)$. Instead of using a hard decision in model selection, we form the estimate for a missing feature using a linear combination of its estimates from all models, where the weights are the *a posteriori* probabilities $p(k|\mathbf{r}(n))$ of models.

In the experiments, we found that good performance is achieved by using both the forward and backward model to reconstruct the final vector by simply averaging the two predictions. In this way, both information from past vectors and future vectors are fully utilized.

Using clean data to train the spectral clusters and their corresponding VAR models gives rise to two types of mismatches. First, the spectral clusters use probabilistic mixture to model the distribution of the clean training data. As a result, the derived

spectral clusters do not describe well distribution of noisy speech data in run-time, leading to inaccurate estimate of cluster *a posteriori* probability $p(k|\mathbf{r}(n))$. The estimate of $p(k|\mathbf{r}(n))$ has direct impact on the quality of reconstructed missing feature. Second, VAR model relies on a sequence of spectral frames to predict a new frame. If the VAR model is trained on clean data, by taking corrupted data in run-time, the VAR model's performance would be unexpected. To address the two mismatches in Scheme I. We propose Scheme II that trains the system using data corrupted by white noise.

5.2.2 Scheme II: Use Noisy Trained VAR Model

In this scheme, noisy data are used to train the system as illustrated in Figure 5.4A. Two approaches are studied, the first uses the noisy data directly to train the system, while the second approach preprocesses the data prior to system training. The training procedure is similar to that of the clean training with two differences. First, the noisy data are used for the spectral clustering; second, the weight matrices B_i are trained by minimizing the prediction error using noisy input vectors to predict the clean desired vector. When reconstructing missing features, the calculation of the *a posteriori* probabilities of VAR models is based on the noisy spectral vectors. The prediction is also based on the noisy spectral vectors of neighboring frames. Therefore, the accuracy of the calculation of the *a posteriori* probability is improved and the mismatch in feature prediction is minimized.

Although the noisy training scheme reduces the mismatches that exists for clean training scheme, there are some inherent technical constraints for this scheme. First, the statistics of the noisy signal changes with signal to noise ratio (SNR). Hence models trained on data of one SNR level are not adequate to reconstruct the noisy speech at another SNR level. Second, for very poor SNR cases (< 5dB), we found that the accuracy of the *a posteriori* probability is low which results in poor reconstruction of features.

To alleviate these problems, the preprocessing module is incorporated into the noisy training and testing processes when the switch chooses preprocessed data $\mathbf{x}'(n)$ (see Fig. 5.4A&B). The objective of the preprocessing module is to condition the noisy speech signal prior to training or testing. Specifically, it aims at reducing the mismatch caused by SNR difference and producing more reliable features. The latter is important, as we assume that for the VAR prediction to be effective, enough reliable features need

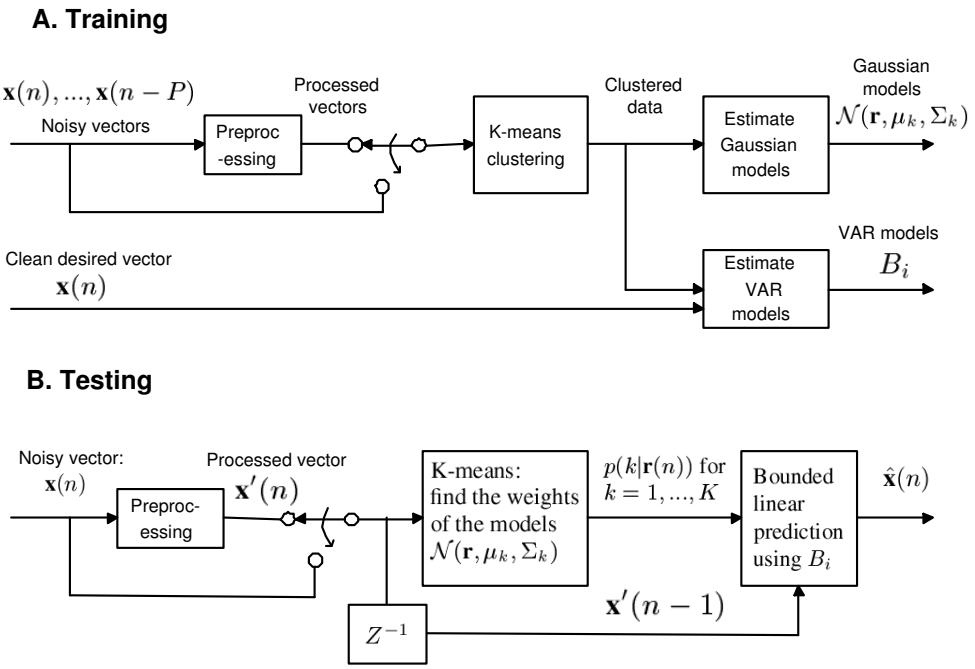

Figure 5.4: Feature compensation scheme II: Use the noisy-trained VAR models with preprocessing

to be present in the input vectors. By making more reliable features, the performance of the VAR prediction will be improved. Many feature compensation and speech enhancement methods may be used for preprocessing, such as Wiener filter and spectral subtraction [17]. In our experiments, the cluster-based feature compensation method [65] is used for preprocessing.

5.3 Experiments on AURORA-2 Database

5.3.1 Setup of the Experiments

The AURORA-2 database [8] was used to evaluate the performance of the proposed feature compensation schemes. The training and testing of the recognition engine follow the scripts provided by the database, except that the c0 is used, rather than log energy. Due to space limits, we only report results on subway noise of test set A and restaurant noise of test set B. As our objective is to examine the performance of the proposed schemes to reconstruct the missing features, we used the oracle binary data mask for our experiments. The optimal SNR threshold is found to be -5dB by experiments [64].

For our two proposed schemes, our experimental results showed that increasing number of clusters improves performance and $C = 50$ cluster is sufficient to model the different classes of speech segments for the AURORA-2 database. We also found $P = 3$ to be a suitable VAR order for the experiments. Hence, these two parameters are used throughout in our experiments as discussed in the following sections.

5.3.2 Results of the Proposed Feature Compensation Schemes

The following six results were obtained for the AURORA Test Set A subway noise as illustrated in Fig 5.5.

a) AURORA-2 baseline model using clean training.

b) Scheme II without preprocessing.

c) Hugo's reported results from [68] using oracle mask.

d) Raj's [64, 65] cluster based MFT method with 20 clusters.

e) Scheme I.

f) Scheme II with preprocessing.

The results showed that our proposed noisy training Scheme II with preprocessing (line-f) gives the best performance with absolute accuracy 88.2% at SNR = -5dB. Compare this result to line-b (noisy training scheme II without preprocessing), the dramatic improvement highlights the importance of the preprocessing steps that conditions the input vectors for the VAR models. Note that the preprocessing step applied for line-f is actually that of Raj's clustering [64, 65] as in line-d. As line-f is significantly better than line-d, this shows that the VAR model further utilizes the inter-frame relationship to improve the reconstruction performance. It implies that the VAR model and Raj's clustering uses complementary information, i.e. the the cluster-based preprocessing module only exploit the intra-frame relationship while the VAR exploits the inter-frame ones. Our experimental results have showed that better result are obtained when these two kinds of information are used together properly.

The results of our proposed clean training Scheme I (line-e) indicates the performance of the VAR-alone scheme. It produces similar performance as Raj's cluster-based method (line-d). This shows that the inter-frame information is as effective as the intra-frame information on the task of missing feature reconstruction.

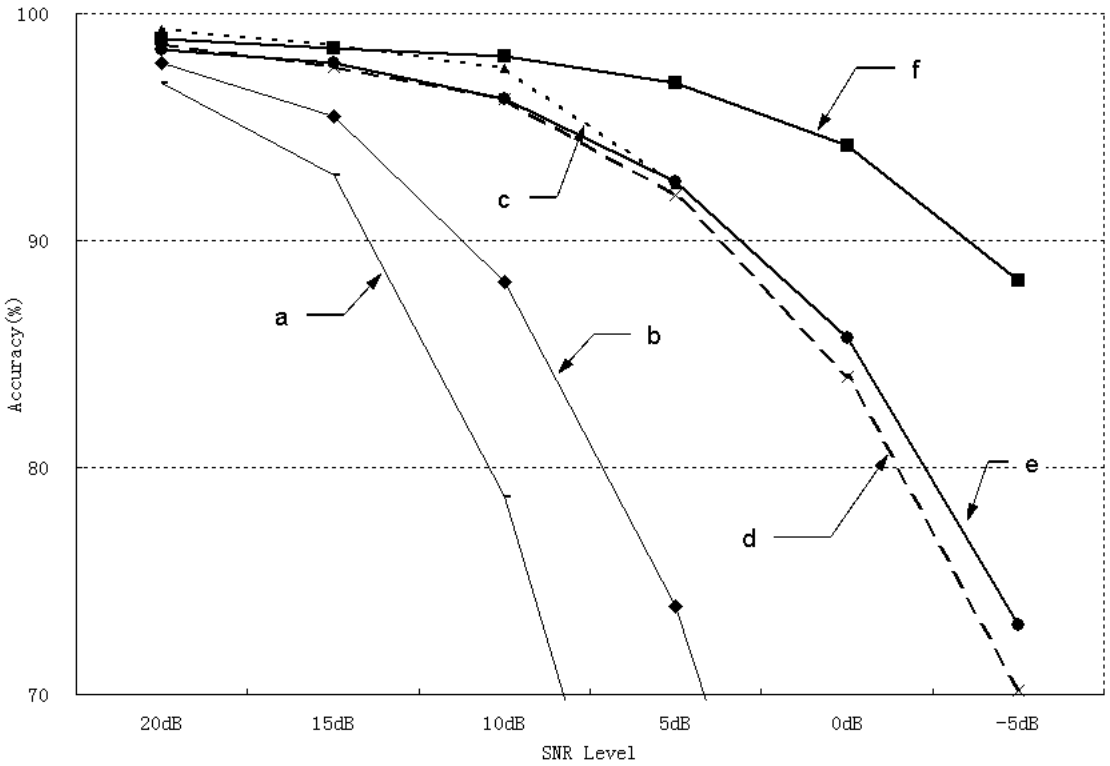

Figure 5.5: Recognition results on subway noise of Test Set A.

The results for the restaurant noise of Test Set B shows similar relative performance as that of subway noise (Fig. 5.6). Here line-a is the AURORA-2 baseline; line-b is for Scheme II without preprocessing; line-c is for Scheme I; line-d is Raj's cluster based MFT reconstruction; and line-e is for Scheme II with preprocessing.

5.4 Summary

In this chapter, we proposed two novel feature compensation schemes using the Vector Autoregressive modeling of spectral vectors. The VAR models are trained on both clean and noisy data respectively. It is found that the models trained on noisy data with the use of preprocessing module provides the best recognition accuracies. The improvement can be credited to the use of both the interframe and intraframe information during feature compensation.

Future research may be conducted to improve the prediction accuracy using nonlinear prediction and use other types of preprocessing techniques. In addition, we will also

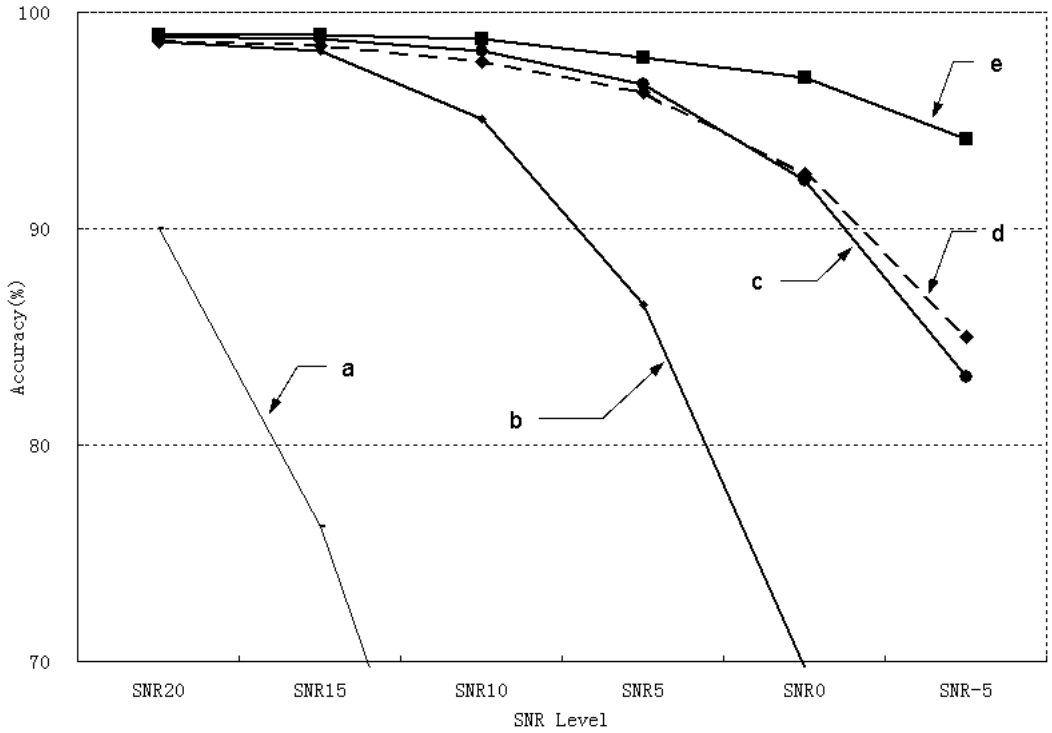

Figure 5.6: Recognition results on restaurant noise of Test Set B.

examine the use of GMM for the spectral clustering. Although our experimental results showed its effectiveness, it is an empirical method and better method for clustering spectral vectors may be used to improve the overall performance. The estimation of VAR model parameters can also be more robust by using advanced model identification methods [15].

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this report, we first reviewed the mismatch problem of the statistical speech recognition due to acoustic environment change and current techniques addressing it. Then the missing feature technique is discussed in detail. To effectively utilize the inter-frame information of the speech spectrogram, we proposed to use the vector autoregressive model for the modeling of speech spectral vectors in the log Mel filterbank domain. Further more, a feature compensation technique is proposed based on this model together with the missing feature theory. The simulation results using oracle data mask showed the effectiveness of the proposed feature compensation technique. Specifically, we compare Raj's MFT-based cluster-based feature compensation technique [65] with our approach . The results show that the two methods are comparable if clean training is used, and much better with our approach when noisy training scheme is used with preprocessing.

In brief, our proposed framework has two novelties. First, we use the vector autoregressive model in the modeling of speech feature vectors. Although VAR is used in [16] for the derivation of the state-dependent probability of the HMM, it has not been used in the feature compensation area to our best knowledge. Second novelty is the use of the noisy training scheme with preprocessing to minimize the mismatches between the training and testing environment.

The potential of the proposed VAR based approach is not fully exploited yet. We believe that further research in this filed will yield fruitful result. Several directions are discussed in the next section.

6.2 Future Works

(i) **Use of estimated mask** Currently, we use the oracle mask to investigate the potential performance of the proposed feature compensation technique. However, to build a workable system, the data mask must be estimated from the noisy observation. We are implementing one of the state-of-the-art mask estimation technique which uses a group of Bayesian classifiers to decide whether a time-frequency location is reliable or not [69]. Once this mask estimation technique is implemented successfully, we plan to test our feature reconstruction techniques based on this mask again. Replacing the oracle mask with an estimated mask, we expect to see recognition accuracy drops. The purpose is to examine the robustness of the system to mask estimation error.

(ii) **Alternative preprocessors** As we discussed before, our proposed noisy training scheme are not constraint to only use the cluster-based technique of [65]. Instead, any speech enhancement and feature compensation techniques could be used as the preprocessor. We will implement other popular speech enhancement and feature compensation techniques as the preprocessor and compare the overall performance. The techniques on our list include spectral subtraction [17]-[20], Ephraim's estimator [21]-[23] and signal subspace techniques [28]-[33].

(iii) **VAR-based mask estimation** It is desirable to apply the VAR model to mask estimation to form a more coherent feature compensation system. Therefore, we will investigate possible methods to apply the VAR model in the mask estimation process.

(iv) **Robustness and non-stationarity** The robustness issue of the estimation of the parameters of the VAR models will be further analyzed. Besides the current least-square based solution, we will also investigate the maximum likelihood estimation and regularized least square approach and compare their robustness [15]. Another important issue is the non-stationary characteristics of speech signal. We believe that current GMM based approach to handle the non-stationary characteristics of speech can be improved. Some example techniques we are going to search is the phoneme based classification of VAR model, the state-space model, etc.

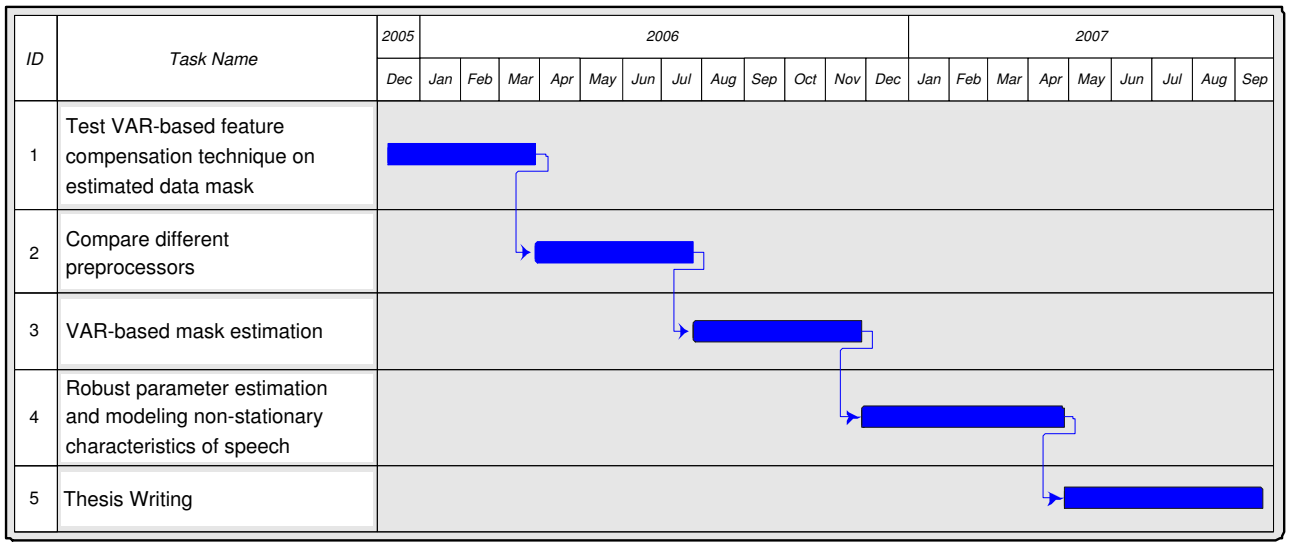

Figure 6.1: PhD research schedule

6.3 Schedule of Future Research

The research plan for the next one year is shown in figure 6.1. Our research will follow this plan, but some variation would be possible due to unexpected findings and difficulties.

Appendix A

Appendix

A.1 Kronecker Product

The Kronecker product is also called matrix direct product and is usually represented as $\otimes$. For example, the Kronecker product a 2×2 matrix A and a 3×2 matrix B is define as

$$\begin{aligned} A \otimes B &= \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix} \\ &= \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{11}b_{31} & a_{11}b_{32} & a_{12}b_{31} & a_{12}b_{32} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \\ a_{21}b_{31} & a_{21}b_{31} & a_{22}b_{31} & a_{22}b_{32} \end{pmatrix} \end{aligned} \tag{A.1}$$

The result is a 6×4 matrix, where any elements from the A is multiplied with any elements from B. For general case, the Kronecker product of a $M \times N$ matrix A and a $P \times Q$ matrix B, the resulting matrix is $MP \times NQ$, and the elements of the resulting matrix is defined as

$$c_{\alpha,\beta} = a_{ij}b_{kl} \tag{A.2}$$

where

$$\alpha = P(i-1) + k \tag{A.3}$$

$$\beta = Q(j-1) + l \tag{A.4}$$

Publication

Submitted

(i) Xiong Xiao, Haizhou Li and Eng Siong Chng, "Vector Autoregressive Model for Missing Feature Reconstruction", submitted to IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2006)

References

[1] R. V. Cox, et. al., "Speech and language processing for next-millennium communications services", Proceedings of the IEEE, Vol. 88, No. 8, Aug. 2000

[2] B. H. Juang, "Speech recognition in adverse environments", Computer Speech and Language, pp. 275-294, Vol. 5, 1991

[3] Y. Gong, "Speech recognition in noisy environments: A survey", Speech communication, pp. 261-291, Vol. 16, 1995

[4] C. H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition", Speech communication, pp. 29-47, Vol. 25, 1998

[5] C. H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition", Proceedings of the IEEE, pp. 1241-1269, Vol. 88, No. 8, Aug. 2000

[6] Y. Ephraim and Isarel Cohen, "Recent Advancements in Speech Enhancement", The Electrical Engineering Handbook, CRC Press, to appear

[7] IBM ViaVoice application

[8] D. Pearce, H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recogntion systems under noisy conditions", *ICSLP, Beijing, China 2000*

[9] S. Young, et. al., "The HTK book", for HTK version 3.2.1, Cambridege university engineering department

[10] L. R. Rabina, "A tutorial on HMM and selected applications in speech recognition", Proceedings of IEEE, pp. 257-286, Vol. 77, No. 2, Feb. 1989

[11] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal", Signal Processing, Vol. 27, No. 1, pp. 65-78, Apr. 1992

[12] M. Ostendorf, V. V. Digalakis, O. A. Kimball, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition", IEEE Trans. on Speech and Audio Processing, pp. 360-378, Vol. 4, No. 5, Sep., 1996

[13] P. Kenny, M. Lennig, P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition", IEEE Trans. on Acoustics, Speech, and Signal Processing, pp. 220-225, Vol. 38, No. 2, Feb., 1990

[14] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of Royal Statistical Society, Vol. 39, pp. 138, 1977

[15] H. Lütkepohl, "Introduction to multiple time series analysis", 2nd Ed., Springer-Verlag, 1993

[16] P. Kenny, M. Lennig and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 38, No. 2, Feb. 1990

[17] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", Proceeding of the IEEE, vol.67, no.12, pp.1586-1604, December 1979

[18] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", in Proc. IEEE ICASSP, pp. 208-211, Apr. 1979

[19] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. Speech and Audio Proc., vol. 7, no.2, Mar.1999

[20] R. J. Mcaulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, no. 2, Apr. 1980

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error Log-spectral amplitude estimator", IEEE Trans. Acoust., Speech, Signal Processing,vol. ASSP-33, pp. 443-445, Apr. 1985.

[23] Olivier Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor", IEEE Trans. Speech and Audio Proc., vol. 3, pp. 251-266, July 1995.

[24] I. Y. Soon and S. N. Koh, "Low distortion speech enhancement", IEE Proc. on Vis. Image Signal Process., vol. 147, no. 3, Jun. 2000

[25] M. K. Hasan et al., "A modified *a priori* SNR for speech enhancement using spectral subtractoin rules", IEEE Signal Processing Letters, vol. 11, no. 4, Apr. 2004

[26] Israel Cohen, "Speech Enhancement Using a Noncausal *A Priori* SNR Estimator", IEEE Signal Processing Letters, vol.11, no. 9, Sep. 2004

[27] Yi Hu and Philipos C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum", IEEE Trans. on Speech Audio Processing, vol. 12, pp. 59-67, Jan. 2004.

[28] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach", Speech Commun., vol. 10, no. 2, pp. 45-57, Feb. 1991.

[29] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Søensen, "Reduction of broadband noise in speech by truncated QSVD",IEEE Trans. Speech Audio Processing, vol. 3, pp. 439-448, Nov. 1995.

[30] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement", IEEE Trans. Speech and Audio Proc., vol. 3, pp. 251-266, July 1995.

[31] J. Huang and Y. Zhao, "An energy-constraind signal subspace method for speech enhancement and recognition in colored noise", in Proc. IEEE ICASSP, vol. 1, (Seattle), pp. 377-380, May 1998.

[32] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement", IEEE Trans. Speech Audio Processing, vol. 9, pp. 87-95, Feb. 2001.

[33] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise", IEEE Trans. Speech Audio Processing, vol. 8, pp. 159-167, Mar. 2000.

[34] B. Dautrich, L. Rabiner, T. Martin, "On the effects of varying filter bank parameters on isolated word recognition", IEEE Trans. on Acoustics, Speech, Signal Processing, pp. 793-806, ASSP-31, Aug. 1983

[35] J. W. Picone, "Signal modeling techniques in speech recognition", Proceedings of the IEEE, Vol. 81, No. 9, Sep. 1993

[36] Y. Epraim and M. Rahim, "On second order statistics and linear estimation of cepstral coefficients", IEEE Trans. on speech and audio processing, pp. 162-176, Vol 7, No. 2, 1999

[37] A. Papoulis and S. U. Pillai, "Probability, random variables and stochastic processes", 4th edition, McGraw Hill, 2002

[38] MathWorld website, available at: "http://mathworld.wolfram.com/"

[39] S. K. Mitra, "Digital signal processing: a computer-based approach", 2nd Edition, McGraw Hill, 2002

[40] T. F. Quatieri, "Discrete-Time Speech Signal Processing: Principles and Practice", Prentice-Hall, 2001

[41] A. V. Oppenheim and R. W. Schafer, with J. R. Buck. "Discrete-Time Signal Processing", 2nd Edition, Prentice-Hall, Upper Saddle River, NJ, 1999

[42] A. V. Oppenheim and R. W. Schafer, "Digital signal processing", Englewood, NJ: Prentice-Hall, 1975

[43] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall, 1993

[44] X. D. Huang, A. Acero, H. W. Hon, "Spoken language processing: A guide to theory, algorithm, and system development", Prentice-Hall 2001

[45] S. M. Kay, "Fundamentals of statistical signal processing: Estimation theory", Prentice-Hall, 1993

[46] G. von Békésy, "Experiments in Hearing", McGraw-Hill, New York, 1960

[47] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern classification", 2nd edition, John Wiley & Sons Inc., 2001

[48] S. Molau, et. al., "Feature space normalization in adverse acoustic conditions", Proceedings of ICASSP 2003, Hong Kong, Apr. 6-10, 2003

[49] F. Liu, et. al., "Efficient cepstral normalization for robust speech recognition", Proceedings of ARPA Human Language Technology Workshop, Mar. 1993

[50] A. Acero, X. Huang, "Augmented cepstral normalization for robust speech recognition", Proceedings of the IEEE Workshop on Automatic Speech Recognition. Snowbird, UT. Dec 1995

[51] P. J. Moreno, B. Raj, R. M. Stern, "Data-driven environmental compensation for speech recognition: A unified appraoch", Speech communication, pp. 267-285, Vol. 24, No. 4, Jul. 1998

[52] L. Deng, J. Droppo and A. Acero, "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior", Proceedings of ICASSP, pp. 829-832, Vol. 1, May 2002

[53] L. Deng, J. Wu, J. Droppo and A. Acero, "Analysis and comparison of two speech feature extraction/compensation algorithms", IEEE Signal Processing Letters, pp. 477-480, Vol. 12, No. 6, Jun. 2005

[54] L. Deng, J. Droppo, A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition", IEEE Trans. on Speech and Audio Processing, pp. 568-580, Vol. 11, No. 6, Nov. 2003

[55] L. Deng, J. Droppo and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features", IEEE Trans. on Speech and Audio processing, pp. 218-223, Vol. 12, No. 3, May 2004

[56] L. Deng, J. Droppo and A. Acero, "Enhancement of Log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise", IEEE Trans. on Speech and Audio processing, pp. 133-143, Vol. 12, No. 2, Mar. 2004

[57] L. Deng, J. Droppo and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametri model of speech distortion", IEEE Trans. on Speech and Audio processing, pp. 412421, Vol. 13, No. 3, May 2005

[58] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition", Speech communication, pp. 231-239, Vol. 12, 1993

[59] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, pp. 171-185, No. 9, 1995

[60] J. L. Gauvain and C. H. Lee, "Maximum *a posterirori* estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Trans. on Speech and Audio Processing, pp. 291-298, Vol. 2, Apr. 1994

[61] Q. Huo, C. Chan and C. H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recgonition", IEEE Trans. on Speech and Audio Processing, pp. 334-345, Vol. 3, Sep. 1995

[62] C. H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition", Speech Communication, pp. 29, 47, Vol. 25, 1998

[63] Q. Huo and C. Lee, "A Bayesian predictive approach to robust speech recognition", IEEE Trans. on Speech and Audio Processing, pp. 200-204, Vol. 8, No. 8, Nov. 2000

[64] B. Raj, R.M. Stern, "Missing-feature approaches in speech reconitoin", *Signal Processing Magazine*, vol. 22, issue 5, pp. 101-116, Sep. 2005

[65] B. Raj, M.L. Seltzer, R.M. Stern, " Reconstruction of missing features for robust speech recognition", *Speech Communication, vol. 43, no. 4, pp. 275-296, Sep. 2004*

[66] M. Cooke, P. Green, L. Josifovski, Vizinho, A., "Robust automatic speech recognition with missing and uncertain acoustic data", *Speech Communication*, vol. 34, pp. 267-285, 2001

[67] Hugo Van hamme, "Robust speech recognition using missing feature theory in the cepstral or LDA domain", Proceedings of EuroSpeech, Geneva, Switzerland, pp. 3089-3092, Sep. 2003

[68] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy mask", *ICASSP 2004*

[69] M. L. Seltzer, B. Raj and R. M. Stern, "A Bayesian framework for spectrographic mask estimation for missing feature speech recognition", Speech Communication, pp. 379-393, Vol. 43, No. 4, 2004

[70] J. Barker, L. Josifovski, M. Cooke and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition", in Proceedings of ICSLP 2000, pp. 373-376, Beijing, China, Sep. 2000

[71] J. Barker, M. Cooke and P. Green, "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise", in Proceedings of Eurospeech 2001, pp. 213-216, Aalborg, Denmark, 2001

[72] K. J. Palomaki, G. J. Brown and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation", Speech Communication, pp. 361-378, Vol. 43, No. 4, Sep. 2004

[73] A. Acero, "Acoustical and environmental robustness in automatic speech recognition", PhD thesis, Dept. of ECE, Carnegie Mellon University, 1990

[74] Y. Ohshima, "Environmental robustness in speech recognition using physiologically-motivated signal processing", PhD thesis, Dept. of ECE, Carnegie Mellon University, 1993

[75] M. J. F. Gales, "Model-based techniques for noise robust speech recognition", PhD thesis, Gonville and Caius College, University of Cambridge, 1995

[76] P. J. Moreno, "Speech recognition in noisy environments", PhD thesis, Dept. of ECE, Carnegie Mellon University, 1996

[77] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition", PhD thesis, Dept. of ECE, Carnegie Mellon University, 2000

[78] S. Molau, "Normalization in the acoustical feature space for improved speech recognition", PhD thesis, Aachen University, 2003

[79] C. P. Chen, "Noise robustness in automatic speech recognition", PhD thesis, Electrical Engineering, University of Washington, 2004

[80] X. Li, "Combination and generation of Parallel feature streams for improved speech recognition", PhD thesis, Dept. of ECE, Carnegie Mellon University, 2005