Speech Enhancement with Applications in Speech Recognition

School of Mechanical and Aerospace Engineering

by

CHEN YONG

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of Doctor of Philosophy

October 15, 2016

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^{2}R) 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^{2}R, 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 Tables

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
diag($\mathbf{x}$)	Make a diagonal matrix whose diagonal elements are the elements of vector $\mathbf{x}$
$\vert\vert\mathbf{x}-\mathbf{y}\vert\vert^2$	Euclidian distance between $\mathbf{x}$ and $\mathbf{y}$
$\vert X\vert$	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) [?] is often used.

Currently, state-of-the-art ASR systems can achieve high recognition accuracy under clean acoustic environment [?]. 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 [?], Wiener filter [?], feature normalization [?, ?] and model-based estimation of the clean speech features [?],[?]-[?]. 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) [?, ?] modifies the distribution of the clean speech to account for the effect of additive noise; maximum likelihood linear regression (MLLR) adaptation [?] techniques transform the means of acoustic model's Gaussians to best fit the noisy observation; maximum *a posteriori* (MAP) [?, ?] adaptation techniques adapt the acoustic model using a Bayesian approach; and STAtistical Reestimation (STAR) [?] 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 [?]-[?]. 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 [?, ?, ?, ?] or discard them during the recognition process [?, ?]. 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 [?] 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 [?], 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 [?]. 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 [?], 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

Current Techniques on Noise Robust ASR

An example to use eps figure:

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. 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.

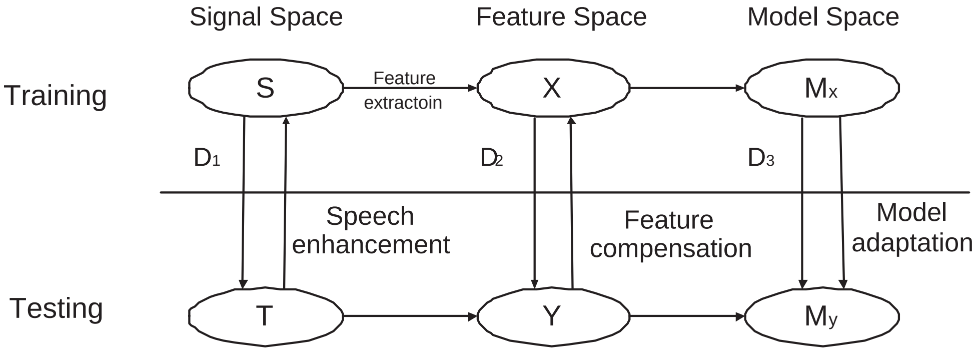

Figure 2.1: The acoustic mismatches in signal, feature and model spaces (adapted from [?]).

Chapter 3

Current Techniques on Noise Robust ASR

Safety and efficiency are two major issues when it comes to vehicle driving and operating, particularly those with relatively large lateral and longitudinal sizes. When performing path planning for vehicles with nonholonomic constraints, we have to consider not only the environment involving obstacles but also the constraints of the vehicle itself. Considering heavy vehicles in a construction site, they have limited maneuverability with a maximum steering angle and are usually maneuvered at a considerably slow speed. To simplify the analysis, the vehicle is supposed to move at a constant speed without slipping thus the mechanical constraints of the vehicle impose a maximum curvature constraint on the reference path. Also, a desired path should have G^2 continuity, i.e., the curve of the path is supposed to have continuous curvature profile to ensure a smooth steering behavior. With respect to a static environment containing narrow corridors, a feasible path should be safe in the sense of maximizing the obstacle clearance while satisfying the above constraints. In this chapter, a new path planning framework for nonholonomic vehicles is proposed to generate a G^2 smooth path with maximum curvature constraint focusing on vehicle driving in a static environment with narrow corridors.

We start with point cloud representation of the environment and obtain a PRM-like map [2] with skeleton algorithm [?]

In this chapter, a new path planning framework for nonholonomic vehicles is proposed to generate a smooth path with maximum curvature constraint. long This chapter discusses a path calculation method for long vehicle turning, based on a set of differential

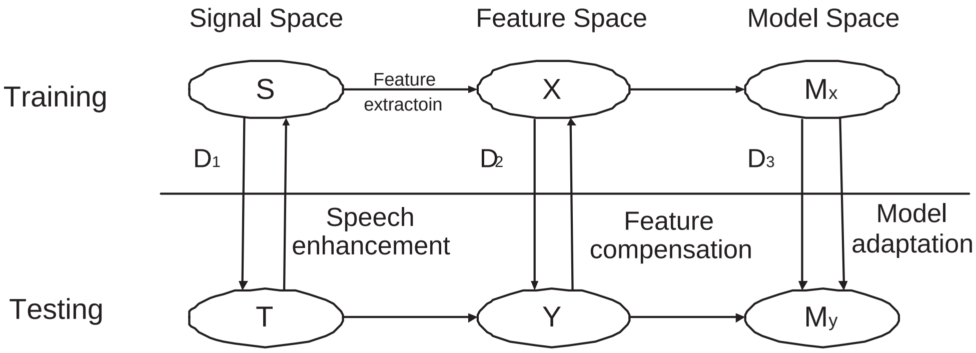

Figure 3.1: The acoustic mismatches in signal, feature and model spaces (adapted from [**?**]).

equations. The solution can be numerically obtained for any trajectory under different circumstances. We develop a generalized and systematic mathematical approach to determine the trajectories swept by each wheel and other related components of the vehicle. The envelope of the trajectories of the vehicle can then be derived according to geometric relationships and characteristics. Based on numerical analysis results, a 3D simulation is developed in this work for different types of long vehicles along with different given turning roads surrounded by buildings and other objects. This way we are able to do trajectory planning for long vehicle turning.

Chapter 4

Conclusions and Future Work

4.1 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 [?] 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 [?] 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.

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

(i) Xiong Xiao, Haizhou Li and Eng Siong Chng, “Vector Autoregressive Model for Missing Feature Reconstruction”, accepted by International Symposium on Chinese Spoken Language Processing 2006.

(ii) Xiong Xiao, Eng Siong Chng and Haizhou Li, “Temporal Structure Normalization of Speech Feature for Robust Speech Recognition”, accepted by to IEEE Signal Processing Letters.

(iii) Xiong Xiao, Eng Siong Chng and Haizhou Li, “Normalizing the Speech Modulation Spectrum for Robust Speech Recognition”, submitted to IEEE International Conference on Acoustics, Speech and Signal Processing 2007.

References

[1] A. Nutbourne, P. McLellan, and R. Kensit, “Curvature profiles for plane curves,” *Computer-aided design*, vol. 4, no. 4, pp. 176–184, 1972.

[2] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars, “Probabilistic roadmaps for path planning in high-dimensional configuration spaces,” *IEEE transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, 1996.