Frequency-Stream-Tying Hidden Markov Model

by

CHONG Fong Ho

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in Department of Electrical and Electronic Engineering

January, 2003, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

CHONG Fong Ho

Frequency-Stream-Tying Hidden Markov Model

by

CHONG Fong Ho

This is to certify that I have examined the above MPhil thesis and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the thesis examination committee have been made.

Dr. Manhung SIU, Thesis Supervisor

Dr. Brian MAK, Thesis Co-Supervisor

Dr. Oscar AU, Thesis Examination Chairman

Dr. James KWOK, Thesis Examination Examiner

Prof. Khaled BEN LETAIEF, Acting Head of Dept. of EEE

Department of Electrical and Electronic Engineering

January 28, 2003

Acknowledgments

Firstly, I would like to thank Dr Manhung Siu and Dr Brian Mak as my thesis supervisor and co-supervisor. Dr Manhung Siu introduced me the concepts and interests on speech technology since I worked on the Final Year Project. He introduced me to Brian Mak, who supported me as a research assistant through the study. Thanks Brian for his patience and helpful guidelines during these years. He not only taught me the concepts of speech technology, but also how to conduct research and present ideas. His encouragement and valuable advice help me a lot for continuing my postgraduate studies.

Secondly, I would like to thank Wilson Tam for giving me advice on my thesis work and allowing me to use his programs on multi-band system. This helped me to save times on my implementation work. I would also like to thank the colleagues in the LASTRE research group for their comments and supports on my work.

Finally, I would like to thank my parents, my brother and my sister. They encouraged me for further studying and forgave me of frequent absence from home during my study period. A special thank should also give to Sara for her support and care.

Contents

Title Page **i**

Authorization Page **ii**

Signature Page **iii**

Acknowledgments **iv**

Table of Contents **v**

List of Figures **xiii**

List of Tables **xix**

Abstract **1**

1 Introduction **3**

1.1 Motivation : Multi-Band Hidden Markov Model 4

1.2 Frequency-Stream-Tying Hidden Markov Model 6

1.3 Organization of this Thesis 8

2 Multi-Band ASR **10**

2.1 Multi-Band ASR 10

2.1.1 Likelihood Recombination Approach 10

2.1.2 Sub-band Asynchrony Issue 12
2.1.3 MCE-derived Sub-band Weights 14
2.2 Other Approaches on Multi-Band System 17

3 Corpus and Baseline Systems 20
3.1 Corpus 20
3.1.1 Training Data 21
3.1.2 Testing Data 22
3.2 Multi-Band Baseline System 23
3.2.1 Equal-Band Partition 23
3.2.2 Multi-Band Baseline Result: Clean-Condition Training 26
3.2.3 Multi-Band Baseline Result: Multi-Condition Training 28
3.2.4 Summary 29
3.3 Full-Band Baseline System 29
3.3.1 How to Choose Full-Band Baseline? 29
3.3.2 Experimental Setup 31
3.3.3 Full-Band Baseline Result: Clean-Condition Training . 32
3.3.4 Full-Band Baseline Result: Multi-Condition Training . 33
3.3.5 Summary 33
3.4 Overall Summary 34

4 Multi-Band System with Optimal Band-partition 37
4.1 Introduction 37
4.2 Experimental Setup 40
4.2.1 Frequency Band Partition Configurations 40
4.2.2 Sub-band Feature Configurations 41
4.2.3 Multi-Band Architecture and Recombination 43
4.2.4 Training and Testing Environment 43

4.3 Experimental Results 43
4.3.1 MB HMM with Optimal Band Partition vs MB HMM with Equal-Band Partition 43
4.3.2 MB HMM with Optimal Band Partition vs FB HMM . 48
4.4 Discussions and Conclusions 49
4.4.1 Discussions 49
4.4.2 Conclusions 51

5 Frequency-Stream-Tying Hidden Markov Model 52
5.1 Motivation 52
5.2 Frequency-Stream-Tying HMM 54
5.2.1 Related Works 54
5.2.2 Topology of the Frequency-Stream-Tying HMM 55
5.2.3 Training Procedures of the Frequency-Stream-Tying HMM 58
5.2.4 Decoding of the Frequency-Stream-Tying HMM 59
5.3 Experiment I: Frequency-Stream-Tying HMM under Clean-Condition Training 60
5.3.1 Experimental Setup 60
5.3.2 Experimental Result 60
5.4 Experiment II: Frequency-Stream-Tying HMM under Multi-Condition Training 65
5.4.1 Motivation : Matched Condition Between the Training and Testing Environments 65
5.4.2 Should Optimal Band Partition still Work in Multi-Condition? 67
5.4.3 Optimization of the sub-band Weights Using Minimum Classification Error Criterion 68

5.5 Experimental III : Full-Band System with Sub-band Energy under Multi-Condition Training 74
5.5.1 Motivation 74
5.5.2 Setup 75
5.5.3 Experimental Result 75
5.5.4 Discussions 79
5.6 Summary 80

6 N-Best Rescoring for Combining the FST System and FB System 83
6.1 Motivation 83
6.1.1 Analysis of the Error Pattern between the FST System and FB System 84
6.2 N-Best Rescoring 86
6.2.1 Methodology 86
6.2.2 Related Works 87
6.3 Experiment 88
6.3.1 Setup 88
6.3.2 Experimental Results 89
6.3.3 Discussions 89
6.4 Summary 93

7 Conclusion and Future Work 94
7.1 Contributions and Conclusions 94
7.1.1 Contributions 94
7.1.2 Conclusions 97
7.2 Future Work 100

Appendices 101

A Aurora Baseline Results **102**

B 16-state Multi-Band HMM vs 18-state Multi-Band HMM **104**

C Detailed Experimental Results **107**

C.1 Clean-Condition Training 107

C.2 Multi-Condition Training 109

D Statistical Significance Tests **112**

References **119**

List of Figures

1.1 A Typical multi-band system 5

1.2 HMM composition of two 3-state left-to-right sub-band HMMs with a maximum asynchrony limit of two states 5

2.1 HMM composition of two 3-state left-to-right sub-band HMMs with a maximum asynchrony limit of two states 13

3.1 3-state full-band HMM vs 3-state multi-band HMM 30

4.1 The frequency spectrum of the four noise signals 39

4.2 The mel-scale filterbank 40

4.3 The recognition results for the multi-band system on car noise under **clean-condition training**, with maximum asynchrony limit up to four states 44

4.4 The recognition results for the multi-band system on subway noise **clean-condition training**, with maximum asynchrony limit up to four states 44

4.5 The recognition results for the multi-band system on babble noise **clean-condition training**, with maximum asynchrony limit up to four states 45

4.6 The recognition results for the multi-band system on exhibition hall noise **clean-condition training**, with maximum asynchrony limit up to four states 45
4.7 Average recognition results for the MB system on the four noises under **clean-condition training**, with maximum asynchrony limit up to four states 47

5.1 A multi-band HMM, with number of states $N_{mb} = 3$, asynchrony limit $A = 2$ 56
5.2 A frequency-stream-tying HMM, with number of states $N_{fst} = 3$, asynchrony limit $A = 2$ 57
5.3 Comparison of the recognition results between the FST system and the FB system on car noise under **clean-condition training** (**FST_w0.5_a1** = FST system with sub-band weights [**0.5:0.5**] and asynchrony limit of one state, **FB 18S** = 18-state FB system) 63
5.4 Comparison of the recognition results between the FST system and the FB system on subway noise under **clean-condition training** 63
5.5 Comparison of the recognition results between the FST system and the FB system on babble noise under **clean-condition training** 64
5.6 Comparison of the recognition results between the FST system and the FB system on exhibition hall noise under **clean-condition training** 64

5.7 Comparison of the recognition results between the FST system with MCE-derived SB weights and the FB system on car noise under **multi-condition training** (**FST_w0.56_a1** = FST system with MCE-derived SB weights [**0.56:0.44**] and asynchrony limit of one state) 71
5.8 Comparison of the recognition results between the FST system with MCE-derived SB weights and the FB system on subway noise under **multi-condition training** 72
5.9 Comparison of the recognition results between the FST system with MCE-derived SB weights and the FB system on babble noise under **multi-condition training** 72
5.10 Comparison of the recognition results between the FST system with MCE-derived SB weights and the FB system on exhibition hall noise under **multi-condition training** 73
5.11 Comparison of the recognition results between the FB system, FB system with SB energy and FST system with MCE-derived SB weights on car noise under **multi-condition training** (**FB_18S** = 18-state FB system, **FB_18S + SBE** = 18-state FB system with SB energy, **FST_w0.56_a1** = FST system with MCE-derived sub-band weights [**0.56:0.44**] and asynchrony limit of one state) 76
5.12 Comparison of the recognition results between the FB system, FB system with SB energy and FST system with MCE-derived SB weights on subway noise under **multi-condition training** 77
5.13 Comparison of the recognition results between the FB system, FB system with SB energy and FST system with MCE-derived SB weights on babble noise under **multi-condition training** 77

5.14 Comparison of the recognition results between the FB system, FB system with SB energy and FST system with MCE-derived SB weights on exhibition hall noise under **multi-condition training** 78

5.15 Summary of the recognition results for the FST system under **multi-condition training**, with the combinations of shifting to optimal band partition (**sub-band partition**), limiting asynchrony of the sub-band transitions (**state asynchrony**) and optimizing the sub-band weights by MCE training (**SB weights**) 80

6.1 Comparison of the recognition results between the merged system using N-Best rescoring and the two individual systems for the car noise under **multi-condition training** 90

6.2 Comparison of the recognition results between the merged system using N-Best rescoring and the two individual systems for the subway noise under **multi-condition training** 91

6.3 Comparison of the recognition results between the merged system using N-Best rescoring and the two individual systems for the babble noise under **multi-condition training** 91

6.4 Comparison of the recognition results between the merged system using N-Best rescoring and the two individual systems for the exhibition hall noise under **multi-condition training** . . 92

B.1 Average recognition results for the 18-state MB system on four noises under **clean-condition training**, with maximum asynchrony limit up to four states 105

List of Tables

3.1 Summary of the Aurora2 training and testing datasets used in this Thesis 22
3.2 Mel-scale filterbank configurations of 22 frequency bands, with the frequency range of 64 Hz to 4000 Hz 24
3.3 Configurations of the equal-band partition baseline system, with **SB Energy** denotes the sub-band Energy 25
3.4 Average word accuracy of the multi-band baseline system under **clean-condition training**, with maximum asynchrony limit ranging from zero to four states 27
3.5 Average word accuracy of the multi-band baseline system under **multi-condition training**, with maximum asynchrony limit ranging from zero to four states 28
3.6 The relationship between number of states N and maximum asynchrony limits L for a multi-band 16-, 18-state HMM . . . 30
3.7 Average word accuracy of the full-band baseline system under **clean-condition training**, varying the number of states per HMM 32
3.8 Average word accuracy of the full-band system under **multi-condition training** 33

3.9 Average word accuracy of the multi-band baseline system and the full-band baseline system under **clean-condition training** (**MB_w0.5_04** = MB system with SB weights [**0.5:0.5**] and asynchrony limit of four states, **FB_16S** = 16-state FB system, **FB_18S** = 18-state FB system) 34
3.10 Average word accuracy of the multi-band baseline system and the full-band baseline system **multi-condition training** . . . 35
3.11 The recognition result for the multi-band baseline system and the full-band baseline system on four **clean** datasets under **clean-condition training** 35
3.12 The recognition result for the multi-band baseline system and the full-band baseline system on four **clean** datasets under **multi-condition training** 36

4.1 Configurations of the different band partitions based on the mel-scale filterbank, where **LLcut**, **LHcut** refer to low pass and high pass cut-off frequency of the first sub-band, **HLcut**, **HHcut** refer to the low pass and high pass cut-off frequency of the second sub-band 41
4.2 Configurations of feature dimensions of different band partitions (**# chans** = number of frequency bands; **# features** = dimension of the acoustic features for the sub-band model; **Total # of features** = total number of acoustic features for the two sub-band models) 42
4.3 Average word accuracy between the optimal band partition multi-band system and the equal-band partition multi-band system 46

4.4 Average word accuracy of the multi-band system with optimal band partition **L-2R+2** setting and the 16-state, 18-state full-band system 49

5.1 The average word accuracy of the frequency-stream-tying system with maximum asynchrony limit ranging from zero to three states under **clean-condition training** 61

5.2 Comparison of the average word accuracy between the FST system, MB system and the FB system on different noises under **clean-condition training** (**MB_w0.5_a2** = MB system with sub-band weights [**0.5:0.5**] and asynchrony limit of two states, **FST_w0.5_a1** = FST system with sub-band weights [**0.5:0.5**] and asynchrony limit of one state, **FST Best** = best result the FST system can achieve by allowing different asynchrony limits on different noises, **FB_16S** = 16-state FB system, **FB_18S** = 18-state FB system) 62

5.3 Comparison of the average word accuracy between the FST system, MB system and the FB system on **clean** dataset under **clean-condition training** 66

5.4 The average word accuracy of the frequency-stream-tying system, with band partition **L-0R+0** and maximum asynchrony limit ranging from zero to three states under **multi-condition training** 67

5.5 The average word accuracy of the frequency-stream-tying system, with band partition **L-2R+2** and maximum asynchrony limit ranging from zero to three states under **multi-condition training** 68

5.6 Comparison of the average word accuracy between the FST system, FST system with MCE-derived sub-band weights, the FB system 70

5.7 The word error rate of the frequency-stream-tying system and the full-band system tested on four **clean** datasets under **multi-condition training**, with the % reduction relative to the 18-state FB system 78

6.1 Comparison of the error pattern similarity e_{sim} for the four noises (**Low SNR datasets**: **5 dB**, **0 dB**, **-5 dB**; **High SNR datasets**: **clean**, **20 dB**, **15 dB**, **10 dB**; **Overall** = overall results over all 7 datasets) 86

6.2 Comparison of the average word accuracy between the merged system using the N-Best rescoring and the two individual systems under **multi-condition training** 90

7.1 Comparison of the average word accuracy (over four noises) between the FST system, MB system and the FB system under **clean-condition training** (**MB/FST** = MB/FST system without optimal band partition and state asynchrony; **FB 16/18S** = FB system with 16/18 states; **OBP** = with optimal band partition **L-2R+2**; **Async** = allow state asynchrony) . . 97

7.2 Comparison of the average word accuracy (over four noises) between the FST system, MB system and the FB system under **multi-condition training** (**FST** = FST system without optimal band partition, without state asynchrony, without MCE-derived sub-band weights; **FB 16/18S** = FB system with 16/18 states; **OBP** = with optimal band partition **L-2R+2**; **Async** = allow state asynchrony; **MCE** = with MCE-derived sub-band weights; **Merged System** = merged system with the use of N-Best rescoring) 98

A.1 Average word accuracy of the Aurora2 baseline system and the full-band baseline system under **clean-condition training** (**Aurora2** = baseline result from the Aurora2 corpus) 103

A.2 Average word accuracy of the Aurora2 baseline system and the full-band baseline system **multi-condition training** . . . 103

B.1 Recognition results for the best 18-state multi-band system and the best 16-state multi-band system on four noises under **clean-condition training** 106

C.1 Recognition results for the best MB system **MB_w0.5_a2** on four noises under **clean-condition training** in Table 7.1 . . . 107

C.2 Recognition results for the best FST **FST_w0.5_a1** system on four noises under **clean-condition training** in Table 7.1 . 108

C.3 Recognition results for the 18-state FB system **FS_18S** on four noises under **clean-condition training** in Table 7.1 . . . 108

C.4 Recognition results for the best MB system **MB_w0.5_a4** on four noises under **multi-condition training** in Table 7.2 . . 109

C.5 Recognition results for the best FST system **FST_w0.56_a1** on four noises under **multi-condition training** in Table 7.2 . 110
C.6 Recognition results for the 16-state FB system **FB_16S** on four noises under **multi-condition training** in Table 7.2 . . 110
C.7 Recognition results for the 18-state FB system **FB_18S** on four noises under **multi-condition training** in Table 7.2 . . 111

D.1 Statistical significance tests on various systems under clean-condition training based on their overall performance on Aurora2 test A (as shown in Table 7.1) 115
D.2 Statistical significance tests on various systems under multi-condition training based on their overall performance on Aurora2 test A (as shown in Table 7.2) 116
D.3 Statistical significance tests on various systems under multi-condition training based on their performance on clean testing data only (as shown in Table 5.7) 117
D.4 Statistical significance tests on various systems under multi-condition training based on their performance on testing data with high SNR: **clean**, **20 dB**, **15 dB** and **10 dB** (as shown in Table C.4 - C.7) 118

Frequency-Stream-Tying Hidden Markov Model

by CHONG Fong Ho

Department of Electrical and Electronic Engineering
The Hong Kong University of Science and Technology

Abstract

In most speech recognition systems, acoustic features are extracted from the whole frequency spectrum of a speech signal. Corruption of the speech signal in one particular frequency band will affect the whole feature vector. It is known that such full-band (FB) speech recognition systems are not robust under noisy environments. To address this problem, multi-band (MB) speech recognition was proposed recently. In the MB approach, the whole frequency spectrum is divided into sub-bands and a model is trained for each sub-band. During recognition, the sub-band results are recombined to yield a global decision. It is hoped that the adverse effect of noises may be mitigated using appropriate recombination logic.

The main contribution of this thesis is a new acoustic modeling method which we call frequency-stream-tying hidden Markov model (FSTHMM). FSTHMM is motivated by the MB HMM generated by HMM composition. An FSTHMM has multiple streams of acoustic features, and each stream consists of acoustic features extracted from one sub-band. These streams are tied across states as in an MB HMM created by HMM composition to allow asynchronous transitions of different sub-band acoustic features. As the number of model parameters remains nearly the same as that of an FB

system, the FSTHMM can be viewed as an effective modeling technique to expand the state space without increasing the number of model parameters. Unlike the MB HMM, the sub-band parameters of an FSTHMM are jointly optimized during training. The sub-band weights can be trained afterward to further improve the system's performance.

In a series of experiments on the Aurora2 corpus, it is shown that FSTHMMs outperform their MB HMM counterparts. The superior performance of the FSTHMMs over the corresponding FB HMMs on clean data but not on noisy data shows that our FSTHMMs should be applied to matched condition only. Using the multi-condition training method, FSTHMMs give a slight improvement over the corresponding FB HMM counterparts on test data with high signal-to-noise ratio (SNR). For data with low SNR, the performance of FSTHMMs varies under different noise environments.

Chapter 1

Introduction

Speech is one of the most natural and direct ways for human communication. It is also the case for the communication between human and computer. For the past several decades, researchers have put a lot of effort developing automatic speech recognition (ASR) technologies with the aim to build a perfect recognizer that can perform speech recognition like humans do.

To build a recognizer that can fulfill the requirement, we should first understand how humans recognize speech. Harvey Fletcher conducted a series of listening experiments in the 1920s and studied how human recognize speech. A set of non-sense consonant-vowel-consonant syllables were low-passed or high-passed for the listening experiments. Based on these experiments, they investigated how humans perform speech recognition on different frequency channels after the contextual information of the speech was removed.

Allen [1] gave a detailed description of Fletcher's study of human speech recognition (HSR). It is believed that partial speech recognition takes place in sub-bands and the sub-band decisions are recombined to give a global decision. It is stated that if two bands are used, the error rate of recognizing

the full band speech e equals to the multiplication of the recognition error rate e_L of low-passed speech and the recognition error rate e_H of high-passed speech.

$$e = e_L * e_H \tag{1.1}$$

The independence between sub-band error rates implies that a human can recognize speech correctly provided that either one of the sub-bands can give the correct answer.

1.1 Motivation : Multi-Band Hidden Markov Model

Based on the understanding of human speech recognition (HSR), researchers have proposed multi-band automatic speech recognition (MBASR) that will try to imitate the HSR process. For a typical multi-band (MB) system, which is shown in Figure 1.1, the speech signal is divided into sub-bands. The acoustic vectors from each sub-band are passed to a sub-band recognizer which will give a sub-band decision. The sub-band decisions will then be combined optimally to reach a final decision.

Figure 1.2 shows an example of the MB HMM. For this MB approach, the sub-band HMMs will be combined to form the MB HMM using the HMM composition algorithm. Under this framework, asynchrony among the sub-band transitions is allowed.

Previous experiments showed that the MB system described above always got better performance than its full-band (FB) counterpart when they are tested on clean speech. However, for artificial noisy data, it can only maintain a similar or degraded performance compared to the corresponding FB system.

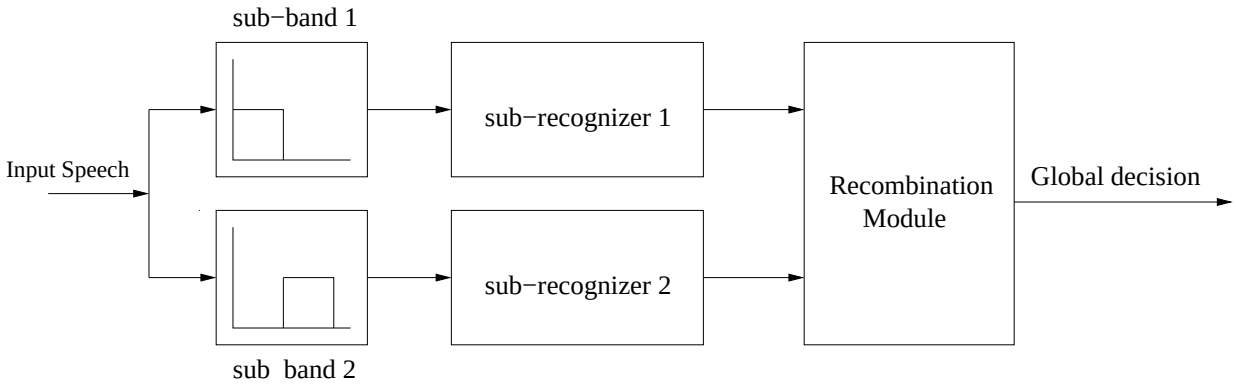

Figure 1.1: A Typical multi-band system

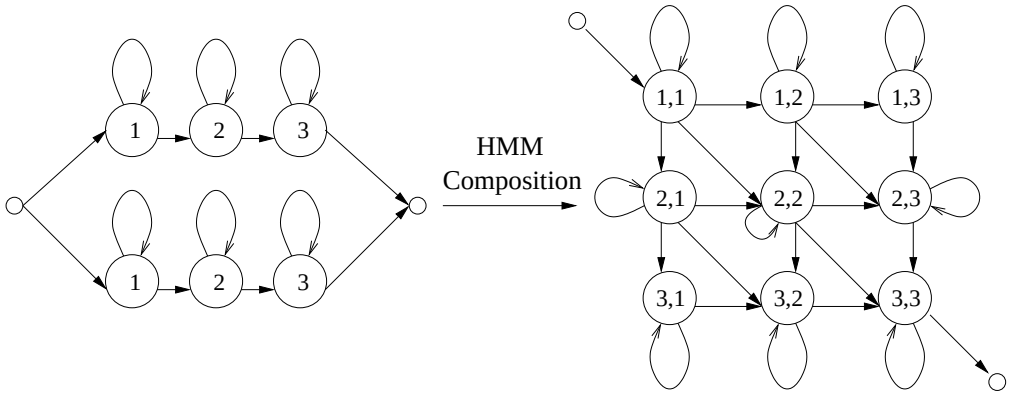

Figure 1.2: HMM composition of two 3-state left-to-right sub-band HMMs with a maximum asynchrony limit of two states

1.2 Frequency-Stream-Tying Hidden Markov Model

In this thesis, we will propose a new acoustic modeling method which we call frequency-stream-tying hidden Markov model (FSTHMM). The FSTHMM is motivated by the MB HMM constructed under the HMM composition framework [2]. It has multiple streams of acoustic features, and each stream consists of acoustic features extracted from one sub-band. The streams of the FSTHMM are tied across the states as in an MB HMM created by HMM composition, which can allow asynchronous transitions of different sub-band acoustic features. Since the new system is also a type of HMM, by applying the traditional training algorithm such as Baum-Welch algorithm, the sub-band parameters of the FSTHMM can be jointly optimized.

For this new proposed frequency-stream-tying HMM, three main points should be highlighted:

1. **Stream tying HMM**

   The topology of the FSTHMM will be exactly the same as that of the MB HMM, which will result in increasing the state space and the state transitions of the FSTHMM, as shown in Figure 1.2. Since the state space of the models are being expanded, the decoding search space and the computation time for the FST system will also increase, compared to the FB system.

   However, as the streams of the FSTHMM are tied across the states as in an MB HMM created by HMM composition, the number of model parameters of the FSTHMM remains nearly the same as that of an FB HMM. The FSTHMM can be viewed as an effective modeling technique to expand the state space without increasing the number of model

parameters.

2. **Asynchronous transitions of sub-band acoustic features**

   Under the FB HMM framework, transitions of sub-band information are forced to be synchronized. Researchers hypothesized that transitions such as phone or state transitions can occur at different times in different bands [3, 4]. This kind of state asynchrony may result in a more accurate modeling of the speech signal.

   By treating the sub-band features as different stream elements of the FSTHMM, the MB model can be viewed as an HMM with multiple streams. Tying is applied on a frequency stream across various states of the model and the frequency-stream-tying HMM can be interpreted as a multi-band HMM. For instance, after tying, a 9-state FSTHMM can be interpreted as a 3-state MB HMM with a maximum asynchrony limit of two states.

   Parameter tying has been applied to many HMM parameters like the mean vectors, variance vectors, Gaussian mixtures, or states. These parameters are usually estimated under the whole frequency spectrum. Unlike these traditional parameter tying scheme, FSTHMM is tied on the frequency stream level. The streams (sub-bands) of different states can be tied together for our FSTHMM and it means that asynchronous recombination of the sub-band information can be allowed at the state level.

3. **Sub-band models are jointly optimized through training**

   Typically, when dealing with sub-band features, researchers will train the sub-band models independently and focus on the recombination of the sub-band information. For this kind of training process, the sub-

band models are optimized independently. The alignments for updating the model parameters during the training process will be independently found and are different for different sub-bands.

However, the various frequency streams FSTHMM, which consists of sub-band model parameters, undergo the training process simultaneously. The sub-band parameters can then be jointly optimized through conventional training algorithm such as the Baum-Welch algorithm. The stream weights corresponding to the weighting of the sub-bands can be trained under the minimum classification error (MCE) criterion afterward to further improve the performance of the FSTHMM.

1.3 Organization of this Thesis

In Chapter 2, a brief description of the multi-band system will be given. Besides, various MB approaches will be included.

In Chapter 3, details of Aurora2, the corpus used to perform experiments in this thesis will be given. Two baseline systems, the full-band baseline system and the multi-band baseline system will also be introduced including their settings and baseline results.

In Chapter 4, we will investigate the performance of the MB model under different band partitions. The hypothesis is that the sub-band partition of the equal-band partition MB model, which separates the acoustic sub-bands such that each band contains the same number of frequency bands, is not an optimal band partition. Under the mel-scale filterbank setting, the full frequency band will be divided into 22 frequency bands. By shifting the band partition left/right by one frequency band each time, a new MB system will be trained and tested to determine the optimal band partition that will give

the best result for the MB system.

In Chapter 5, we will concentrate on the frequency-stream-tying HMM, the new proposed HMM motivated by the composite HMM. Through the tying of the frequency streams, the new system can allow asynchrony in sub-band transitions, which is the same as the MB system. The FSTHMM will partition speech based on the optimal band partition found in Chapter 5. A set of FSTHMMs with the maximum asynchrony limit up to 3 states will be trained. To further enhance the system, the sub-band (streams) weights are trained using the MCE criterion.

In Chapter 6, we will study the possibilities of combining the FB system and the FST system to enhance recognition performance by N-Best rescoring. The better system will be chosen as the first system to generate the N-Best sequences, while the second system will be used to rescore the N-Best sequences. The experimental results of the combined systems will be given at the end of this chapter.

In Chapter 7, a summary of the contributions of this thesis will be given. Further improvements of this frequency-stream-tying framework will also be proposed.

Chapter 2

Multi-Band ASR

In this chapter, the detailed description of our multi-band system will be given. We will mainly concentrate on three areas: the likelihood recombination approach of the sub-bands, the asynchrony issue between sub-band transitions under the HMM composition framework and the optimization of the sub-band weights under the minimum classification error (MCE) criterion. Besides the approach used in our MB system, other MB approaches will be discussed.

2.1 Multi-Band ASR

2.1.1 Likelihood Recombination Approach

This approach was first proposed by Bourlard *et al.* [5]. Among the current approaches of recombining sub-bands, the likelihood recombination approach has been commonly explored by researchers in recent years. Other recombination approaches beyond the likelihood recombination are also proposed. Hermansky *et al.* [6] have used the majority voting scheme to combine decisions from the individual sub-band combination. Cerisara *et al.* [7] have

recombined the phoneme sequences proposed by each sub-band on the task of phoneme recognition.

To present the mathematical formulation of the likelihood recombination approach, a series of definitions are given first. Let x be a full-band acoustic vector sequence; and M be the acoustic model (word or sentence); X_j is be the j-th segment of X; and M_j is be the model associated with X_j. The log likelihood of x is then be defined as

$$\log P(X|M) = \max_{J,M_j} \sum_{j=1}^{J} \log P(X_j|M_j) \tag{2.1}$$

The recombination of the sub-band is further divided into linear recombination or non-linear recombination strategies. By the use of the linear recombination strategy, the sub-band log-likelihoods are simply the weighted sum for each segment of speech, as shown in the following equation:

$$\log P(X_j|M_j) = \sum_{k=1}^{K} w_k \log P(X_j^k|M_j^k) \tag{2.2}$$

where K is the total number of sub-bands; w_k is the k-th sub-band weights ($k = 1, \ldots, K$). X_j^k is the j-th segment of X corresponding to the k-th sub-band and M_j^k is the model associated with X_j^k. The sub-band weights can be treated as a reliability measure by setting the more reliable sub-band a larger weight.

For simplicity, the sub-band weights is set as equal in this thesis to assume the equal contribution of the sub-bands. Beside the equal weighting scheme, other schemes have also been proposed to estimate the sub-band weights. Hermansky *et al.* [6] have tried the sub-band weights which is estimated by the accuracy of the individual sub-bands. Okawa *et al.* [8] have used the sub-band signal-to-noise ratio (SNR) computed at frame level as the weighting scheme of the sub-band likelihood recombination.

Other than the sub-band SNR weighting scheme, Okawa *et al.* [8] have also recombined the sub-band likelihood by the use of entropy weighting. The sub-band weights are equal to the inverse of the conditional entropy, which is computed from the a posteriori probabilities of all HMM states. The weights are normalized to sum up to the number of sub-bands. Experimental results have shown that the MB system can achieve a significant improvement over the FB system when testing on speech corrupted by the low-pass white noise. Besides, Okawa *et al.* [9] have also used the mutual information weighting scheme derived from the information theory for estimating the sub-band weights.

The non-linear recombination strategy for the sub-band likelihood can be achieved by the use of multiple layer perceptron (MLP) [5, 6, 10, 11]. A brief description on this approach is given in the next section.

2.1.2 Sub-band Asynchrony Issue

Researchers have hypothesized that the transitions of sub-band can occur at different times in different bands and, therefore, it may take the advantage to perform asynchronous recombination of sub-bands. The MB system used in this thesis will makes use of the HMM composition framework to incorporate asynchrony between the sub-bands.

The HMM composition is originally applied to address the problem of speech recognition in the presence of noise [12]. Similar to the Parallel Model Combination (PMC), the contaminated speech can be thought of containing two separate components (speech signal and noise). Since the two individual components can be modeled by conventional HMMs, the signal results from the combination of speech and noise can be modeled as a function of their combined outputs. In other words, recognition of the speech is carried out

with one set of model while the noise is "recognized" simultaneously by the other set of models. Mirghafori *et al.* [13] have applied the HMM composition method in MBASR in which the speech model and the noise model are replaced with the sub-band HMMs. Through the HMM composition, the sub-band models are combined to form the composite HMM.

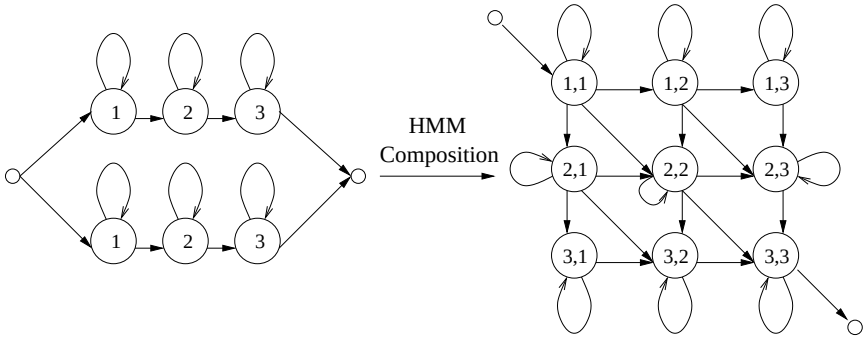

Figure 2.1: HMM composition of two 3-state left-to-right sub-band HMMs with a maximum asynchrony limit of two states

Figure 2.1 illustrates the idea of the HMM composition algorithm. The two-band system consisting of two 3-state left-to-right sub-band HMMs are converted to the 3-state MB HMM with a maximum asynchrony of two state. As shown in the figure, the composite HMM has 9 "composite states" $S_{i,j}$, for $i, j = 1, \ldots, 3$, where state i of the first sub-band HMM recombines with state j of the second sub-band HMM. For instance, the composite state $S_{1,3}$ means that the first state of the first sub-band HMM recombines with the third state of the second sub-band HMM. The composite state $S_{1,3}$ also indicates that the sub-bands are in different HMM states, with an asynchrony of two states. In general, a synchronous state can be represented by $S_{i,i}$ while a asynchronous state can be represented by $S_{i,j}$, for $i \neq j$.

In general, for a K-band system, lets define $L_{(i_1,\ldots,i_K)}$ as the log likelihood of a composite state $(i_1,\ldots,i_K)$ and L_{ik} as the log likelihood of state i_k in sub-band k. By the use of linear combination strategy, the log likelihood of the composite state is the weighted sum of the log likelihoods of its corresponding sub-band states:

$$L_{(i_1,\ldots,i_K)} = \sum_{k=1}^{K} w_k L_{i_k} \tag{2.3}$$

where w_k denotes the weighting of the k-th sub-band with the constraint that the sum of sub-band weights equals to 1.

The transition probabilities of the composite HMM is then computed as follows:

$$a_{(i_1,\ldots,i_K),(j_1,\ldots,j_K)} = \frac{\sum_{k=1}^{K} a_{i_k,j_k}}{\sum_{s\in S} a_{(i_1,\ldots,i_K),s}} \tag{2.4}$$

where a_{i_k,j_k} is the transition probability from state i_k to j_k in sub-band k. The denominator is used to normalized the transitional probabilities where S is the set of states that $(i_1,\ldots,i_K)$ transits to [2].

Other than the HMM composition framework, Cerisara *et al.* [14] have proposed an efficient adaptation of the two-level DP algorithm and the level building DP to allow the bands to be asynchronous in their MB system. The new algorithm has shown that continuous MB speech recognition can be achieved at low computational cost. It is stated that approximately one third of sub-band frames were asynchronous according to the sub-band alignments. The accuracy of the MB system with the incorporation of asynchrony can performs at least as well as the synchronous MB system.

2.1.3 MCE-derived Sub-band Weights

The sub-band weights of the MB system can be "optimized" based on the MCE criterion [15, 16]. MCE training is a common discriminative technique

used to optimize system parameters by minimizing the system's classification errors. MCE training involves two steps. The misclassification measure $d(X)$ is defined first and a sigmoid function $l(d(X))$ is then applied to evaluate the "soft" error. The misclassification measure is defined as:

$$d(X) = \log \left\{ \frac{1}{N} \sum_{n=1}^{N} e^{g_n \eta} \right\}^{1/\eta} - g_0 \tag{2.5}$$

where $g_j, (j = 0, \ldots, N)$ denotes a set of N discriminant functions for each of the N competing hypotheses. g_j is the log likelihood of the j-th competing hypothesis. η is the parameter to control the participation of all competing hypotheses and it is set as 1 for equal participation of the competing hypotheses. $d(X)$ can be viewed as the "average" distance of competing hypotheses from the correct hypothesis for an utterance X.

The misclassification distance $d(X)$ is smoothed using the sigmoid function $l(d(X))$:

$$l(d(X)) = \frac{1}{1 + e^{-\gamma(d(X))+\theta)}} \tag{2.6}$$

The sigmoid function contains two parameters namely γ and θ. γ controls the slope of the sigmoid function while θ determines the offset of the sigmoid function along the horizontal axis. The two parameters can affect the efficiency and effectiveness of MCE training. The setting of these parameters should be carefully tuned because making the slope of the sigmoid too steep will leads to little training samples involved. However, if the slope is too flat, the training process will be small.

To perform MCE training on continuous speech recognition, the competing hypotheses against the correct hypothesis are commonly derived from the N-Best algorithm. However, some of the misclassified data have a large misclassification distance from the N-Best competitors, resulting in a fall-out from the steep/trainable region of the sigmoid function. An alternative

approach is to use the nearest competing hypothesis instead of the N-Best hypotheses for MCE training [17]. The latter approach can keep the training data as close to the training region as possible. It helps to increase the amount of "effective" training data because more training data are falling within the trainable region of the sigmoid function.

For a K sub-band recognition system and N competing strings (hypotheses). $X = [x_1, x_2, \ldots, x_T]$ denotes the observation sequence and Λ denotes the acoustic model sets. The discriminant function for class W_j will be

$$g_j(X : \Lambda) = \log P(X, \overline{q}_j | W_j) \tag{2.7}$$

where $\overline{q}_j = \{\overline{q}_j^1, \overline{q}_j^2, \ldots, \overline{q}_j^K\}$ denotes the most likely composite state sequence of W_j and $\overline{q}_j^k$ corresponds to the state sequence in sub-band k. The sub-band log-likelihoods are recombined at the frame level as shown below:

$$\log b_{\overline{q}_j(t)}(x_t) = \sum_{k=1}^{K} w_k \log b^k_{\overline{q}_j(t)^k}(x_t^k) \tag{2.8}$$

with the constraints $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$.

Given the training dataset X_i $(i = 1, \ldots, U)$, the empirical loss function for minimization is given by

$$R_{mce} = \frac{1}{U} \sum_{i}^{U} l(d(Xi)) \tag{2.9}$$

To ensure the positivity of the sub-band weights w_k throughout the optimization process, the following transformation pairs are applied:

$$w_k = \frac{e^{\overline{w}_k}}{\sum_{k=1}^{K} \overline{w}_k} \tag{2.10}$$

$$\overline{w}_k = \log(w_k) \tag{2.11}$$

Taking derivative with respect to the k-th sub-band weight $\overline{w}_k$, we have

$$\frac{\partial R_{mce}}{\partial \overline{w}_k} = \frac{1}{U} \sum_{x} \frac{\partial l}{\partial d} \sum_{b}^{K} \frac{\partial d(X)}{\partial w_b} \frac{\partial w_b}{\partial \overline{w}_k}, \text{ where} \tag{2.12}$$

$$\frac{\partial l}{\partial d} = \gamma l(d(X))\,[1 - l(d(X))] \tag{2.13}$$

$$\frac{\partial d(X)}{\partial w_b} = \frac{\sum_{n \neq j}^{N} exp(g_n)(\frac{\partial g_n}{\partial w_b} - \frac{\partial g_j}{\partial w_b})}{\sum_{n \neq j}^{N} exp(g_n)} \tag{2.14}$$

$$\text{where} \quad \frac{\partial g_n}{\partial w_b} = \sum_{t=1}^{T} \log b^b_{\overline{q}_n(t)}(x^b_t) \tag{2.15}$$

The gradient-descent algorithm can be iteratively used until the empirical loss function is minimized. The sub-band weights are then optimized under the MCE criterion.

$$\overline{w}_k^{(t+1)} = \overline{w}_k^{(t)} - \epsilon_t \left(\frac{\partial R_{mce}}{\partial \overline{w}_k} \right) \tag{2.16}$$

where ϵ_t is the learning rate of the t-th iteration, and the estimate of the sub-band weight for $(t+1)$-th iteration depends on the estimate of the t-th iteration.

2.2 Other Approaches on Multi-Band System

The MB systems described in the previous section recombine the sub-band log-likelihoods based on the linear recombination strategy. Researchers believe that non-linear recombination of sub-band information occurred in humans. The non-linear recombination of sub-band likelihoods can be achieved by the use of multiple layer perception (MLP) [5, 6, 10, 11]. The MLP contains an input layer, hidden layers and an output layer. For an MB system with M word models and each with K sub-bands and suppose the sub-band log-likelihoods are recombined at the word level using the MLP, the $M * K$ inputs will be fed into the MLP to estimate the M outputs. The weights between the three different layers are trained using the back-propagation algorithm.

For simplicity, the MB system usually consists of two sub-bands. Researchers have built MB systems with 3 to 7 sub-bands to investigate the effectiveness of different merging policies [5, 6, 10]. Hermansky *et al.* [6] have built a 7 sub-band MB system in which the nonlinear merging networks are trained using all total 127 set ($2^7 - 1 = 127$) of possible sub-band combinations. Experimental results have shown that there exist at least one sub-band combination that can yield a very good result in the present of selective noise. It is reported that linear weighting schemes appear to be feasible for 2 sub-band approach but not for the 7 sub-band approach. Although the MLP merging scheme can achieve the best performance for both two systems, the 2 sub-band system performs better.

Okawa *et al.* [8] have developed an MB system with the recombination of sub-bands are held at feature level. After the cepstral feature for each sub-band is extracted individually, they are combined into one single vector. The whole feature vector is then fed as the input of the trainer/recognizer. Such kind of approach is believed to be possible for modeling the correlation between the sub-bands and can avoid the complicated weighting strategies for sub-band recombination.

Other than the recombination of sub-band likelihoods, Cerisara *et al.* [7] have developed an MB system to merge the label of each sub-band. The phoneme sequences proposed by the individual sub-band are recombined to form a global phoneme sequence.

McCout *et al.* [18] have proposed the use of multi-resolution sub-band cepstral features to perform phoneme recognition. The multi-resolution cepstral features are combined with the various multi-resolution of sub-band features and the full-band cepstral features. The individual log likelihoods are summed at the frame level. The best results are given by combining

the full-band system with either a two-band or a four-band system for the TIMIT phone-recognition task.

Hagen *et al.* [19] have proposed the Full Combination (FC) multi-band system in which the MLP merging scheme have been replaced by the Multi-Gaussian HMM (MGM) in the full-combination of sub-bands. By applying Bayes' Rule, the likelihood-based FC formula can be easily derived by the posterior-based MLP.

Chapter 3

Corpus and Baseline Systems

In this chapter, the corpus used to perform recognition experiments in our research is introduced. Two baseline systems: the multi-band baseline system and the full-band baseline system will be described. The latter part of this chapter will give the reasons for the baseline configurations and the baseline results.

3.1 Corpus

The corpus used to perform the experiments in this thesis is the Aurora2 corpus [20]. It is a database developed by the AURORA DSR (Distributed Speech Recognition) Working Group of the ETSI STQ [21] and designed for evaluating the performance of speech recognition algorithms in noisy conditions. The source speech for the Aurora2 database is the TIDigits [22], which contains recordings of male and female American adults speaking connected digits, with a maximum of 7 digits in a sequence. The original 20kHz data are first down sampled to 8kHz with an ideal low pass filter extracting the

spectrum between 0 to 4kHz. An additional filtering (G.712 or MIRS defined by the ITU) is then applied to consider the realistic frequency characteristics of terminals and equipment in the telecommunication area. A selection of 8 different real world noises have been added to the speech over a range of signal to noise ratios (SNR). In my thesis, only part of the dataset is used for performing the experiments.

3.1.1 Training Data

Two training modes are defined in the Aurora2 corpus: training on clean (**clean-condition**) data only, and training on clean and noisy (**multi-condition**) data on various SNR.

Clean-Condition Training (CleanCond)

The training set consists of 8440 utterances from 55 male and 55 female adult speakers. The speech data are filtered with the G.712 without adding noises.

Multi-Condition Training (MultiCond)

The training set consists of the same 8440 utterances in clean-condition training dataset. The training data are equally split into 20 subsets so that each subset contains a few utterances of all training speakers. The 20 subsets represent 4 different noise scenarios (suburban train, babble, car, exhibition hall) at 5 different SNRs of **clean**, **20 dB**, **15 dB**, **10 dB**, **5 dB** conditions. Before adding the noises, both the speech and noise data are filtered with the G.712.

3.1.2 Testing Data

In this thesis, we used the Aurora2 Test A only. The testing data consists of 4004 utterances from 52 male and 52 female speakers in TIDigits testing data. The data are split into 4 subsets with 1001 utterances in each subset. Four types of noises in **test set A** (suburban train, babble, car, exhibition hall) are added to the 4 subsets at 7 different SNRs of **clean**, **20 dB**, **15 dB**, **10 dB**, **5 dB**, **0 dB** and **-5 dB** conditions. In total, the testing set consists of $4*7*1001 = 28028$ utterances. Except for the **0 dB** and **-5 dB** testing data, the test set matches the acoustic condition of the multi-condition training dataset.

Table 3.1 shows a summary of the two training datasets and testing dataset of the Aurora2 that was used in our experiments.

Table 3.1: Summary of the Aurora2 training and testing datasets used in this Thesis

	Training		
	CleanCond	**MultiCond**	**Testing**
# Subsets	1	20	28
# Noise Types	0	4	4
# Speakers	55 M, 55 F	55 M, 55 F	52 M, 52 F
# Utterances / Subset	8440	422	1001
# Total # Utterances	8440	8440	28028

3.2 Multi-Band Baseline System

In this thesis, a two-band system was used for all experiments. The acoustic features used in this thesis are Mel-Frequency Cepstral Coefficients (MFCC). Further details for generating the MFCCs are:

- Analysis window size of 25 ms using the Hamming window
- Frame shifting rate of 10 ms
- Zero-mean the speech data
- Enable energy normalization for energy computation
- Use of delta and acceleration cepstral coefficients
- FFT based Mel filterbank with 22 frequency bands in the range from 64 Hz to 4000 Hz for combining the two sub-band systems

3.2.1 Equal-Band Partition

Based on the Mel-frequency filterbank scale [23], the full frequency was partitioned into 22 frequency bands. Table 3.2 shows the Mel-scale filterbank configurations of 22 frequency bands, with the frequency range of 64 Hz to 4000 Hz.

The baseline system then partitioned the frequency into sub-bands so that each band contains the same number of frequency bands and this is referred to the equal-band partition system. The sub-band (SB) energy is calculated in the frequency spectrum domain under the frequency partition specified by the individual sub-band. Table 3.3 shows the details of the sub-band configurations.

Table 3.2: Mel-scale filterbank configurations of 22 frequency bands, with the frequency range of 64 Hz to 4000 Hz

Frequency Band Index	**Frequency (Hz)**		
	start	**center**	**end**
1	64	127	195
2	127	195	268
3	195	268	348
4	268	348	434
5	348	434	527
6	434	527	628
7	527	628	737
8	628	737	855
9	737	855	983
10	855	983	1121
11	983	1121	1271
12	1121	1271	1433
13	1271	1433	1609
14	1433	1609	1798
15	1609	1798	2004
16	1798	2004	2226
17	2004	2226	2466
18	2226	2466	2727
19	2466	2727	3008
20	2727	3008	3313
21	3008	3313	3643
22	3313	3643	4000

Table 3.3: Configurations of the equal-band partition baseline system, with **SB Energy** denotes the sub-band Energy

	sub-band 1	**sub-band 2**
Frequency Range (Hz)	64 - 1271	1121 - 4000
Number of Frequency Bands	11	11
Static Acoustic Features	6 MFCC + SB Energy	6 MFCC + SB Energy

HMM Architecture

The individual SB models are trained independently. The 11 digits ("zero" – "nine", "oh") are modeled as whole word HMMs. Each HMM contains 16 real states and are strictly left-to-right. For each state, there are three Gaussian mixtures. Besides the digits, there are two silence models namely "sp" and "sil". The "sil" HMM consists of three real states while the "sp" HMM consists of one real state. The architecture of the two silence models have the same setting as the Aurora2 baseline system [20].

Sub-band Recombination

The SB models are combined into an MB model using the HMM composition algorithm described in Section 2.1.2. The log likelihood of the composite state is computed by the weighted sum of the sub-band likelihoods with a weighting scheme of [**0.5:0.5**]. A maximum asynchrony limits of zero to four states are allowed for the MB system.

Training and Testing Environment

The MB HMMs are trained on the clean-condition training data and multi-condition training data. The models are evaluated on the testing set which contains the four noises: car, subway, babble and exhibition hall noise. The word insertion penalty is individually tuned on the MB system when testing on low SNR datasets so the MB system achieves its best result. The word insertion penalty is not tuned when testing on all testing sets is because we find that the number of insertions and the number of deletions during recognition are not symmetric for the low SNR case. However, it is not the case for the high SNR dataset, tuning the word insertion penalty will harm the performance.

3.2.2 Multi-Band Baseline Result: Clean-Condition Training

For the Aurora2 testing data, there are 28 subsets, 4 different type of noises with 7 SNR conditions in total. It is impossible to show all the MB baseline results which contain 5 different asynchrony limits. For simplicity, for each noise, only the average result will be shown. The average result will only include the results of **0 dB** dataset to **20 dB** dataset (a total of 5 SNR conditions), which follow the setting of the Aurora2 experiments. Table 3.4 shows the average performance (word accuracy) of the MB baseline system under clean-condition training, with the maximum asynchrony limits ranging from zero to four states (referred as a0, a1, a2, a3, a4, respectively).

From the results tabulated in Table 3.4, we found that:

- The asynchronous MB system always performed better than the synchronous MB system. The average word error rate is reduced by about

Table 3.4: Average word accuracy of the multi-band baseline system under **clean-condition training**, with maximum asynchrony limit ranging from zero to four states

Noise Type	Maximum Asynchrony Limits				
	a0	**a1**	**a2**	**a3**	**a4**
Car	51.05	54.46	55.3	54.75	54.89
Subway	55.01	57.32	59.11	58.27	59.89
Babble	51.49	54.35	55.46	55.69	55.93
Exhibition	50.17	53.41	54.26	53.91	53.95
Average	51.93	54.89	56.03	55.66	56.17

3% absolutely comparing to the fully synchronous system (with maximum asynchrony limit $L = 0$) and the asynchronous system (with maximum asynchrony limit $L = 1$).

- When the maximum asynchrony limit L is increased to two, there is still improvement of the average word accuracy. However, further increasing L will not help improve the MB system significantly. Since similar results are obtained when L is greater than two, it seems the performance has flattened at the asynchrony limit of two.

- Limiting asynchrony to a maximum of four states will usually give the best result of the MB baseline system.

Table 3.5: Average word accuracy of the multi-band baseline system under **multi-condition training**, with maximum asynchrony limit ranging from zero to four states

Noise Type	Maximum Asynchrony Limits				
	a0	**a1**	**a2**	**a3**	**a4**
Car	83.78	87.12	88.27	88.69	88.75
Subway	82.11	86.28	87.29	87.90	88.00
Babble	82.74	84.90	85.92	86.40	86.51
Exhibition	80.51	83.86	85.07	85.70	85.80
Average	82.28	85.54	86.64	87.17	87.27

3.2.3 Multi-Band Baseline Result: Multi-Condition Training

Table 3.5 shows the average performance (word accuracy) of the MB baseline system, with the maximum asynchrony limits ranging from zero to four states.

From the results tabulated in Table 3.5, we found that:

- The asynchronous MB system always performed better than the synchronous MB system. The average word error rate is reduced by about 3% absolutely comparing to the fully synchronous system (with maximum asynchrony limit $L = 0$) and the asynchronous system (with maximum asynchrony limit $L = 1$).

- Increasing the maximum asynchrony limit L will always help to improve the average word accuracy. The best result of the MB baseline system can be achieved by limiting the maximum asynchrony to four states.

3.2.4 Summary

For both the clean-condition training and multi-condition training, HMM composition is an effective way to implement asynchronous recombination of sub-bands in multi-band ASR. Through this framework, SB models are recombined asynchronously. By allowing a asynchrony limit of 1 state, it helps to reduced the average word error rates for the Aurora2 testing dataset by about 3%.

For both the clean-condition training and multi-condition training MB baseline system, the best results can be achieved by limiting asynchrony to a maximum of four states. It is therefore chosen as the baseline results.

3.3 Full-Band Baseline System

Besides comparing to the MB baseline system, the new proposed frequency-stream-tying system should also be compared with the full-band system to see if the it can outperform the FB system or not.

3.3.1 How to Choose Full-Band Baseline?

Previous research showed that under the noisy environments, an asynchronous MB system can yield better performance compared with the full-band system [24]. However, some may argue that after allowing asynchrony, the state space of the MB system is expanded. Although the number of parameters of the MB system still remain the same, the number of states per each HMM is increasing according to the maximum asynchrony limit of the MB system, as shown in Figure 3.1.

Table 3.6 shows the relationship of the number of states with maximum asynchrony limits up to 4 (referred as a0, a1, a2, a3, a4, respectively) for

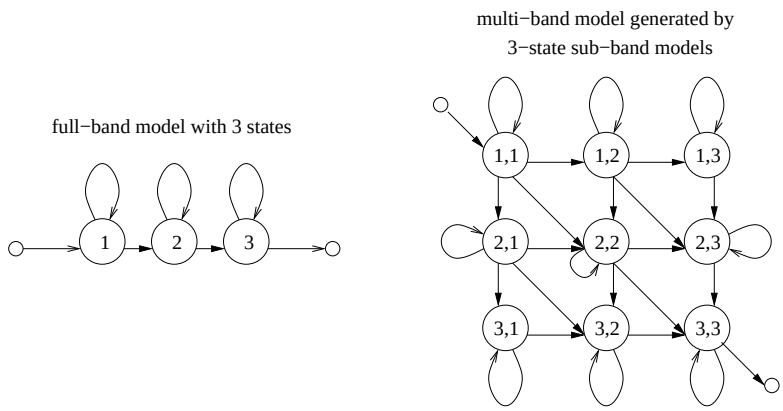

Figure 3.1: 3-state full-band HMM vs 3-state multi-band HMM

both MB 16-state and 18-state HMM.

Table 3.6: The relationship between number of states N and maximum asynchrony limits L for a multi-band 16-, 18-state HMM

Number of States Per HMM					
Full-Band System	**Multi-Band System**				
	Maximum Asynchrony Limit				
	a0	**a1**	**a2**	**a3**	**a4**
16	16	46	74	100	124
18	18	52	84	114	142

The table shows the number of states of the MB HMM is much larger than that of the FB system. Researchers thus argue that the MB system performs better than the FB system just because of the expansion of the state space, which may lead to a better resolution for modeling the speech, but not because of the characteristics and the nature of the MB system.

In order to get a fair comparison between the two systems, we are going to increase the number of the states of the HMM for the FB system. Through

the experiments, we will choose the one with the best result to be the full-band baseline.

3.3.2 Experimental Setup

Configurations of the full-band system

The process of feature extraction is the same as that of MB baseline system except:

- Frequency partition : 64 - 4000 Hz
- Number of frequency bands : 22
- Number of MFCCs : 12 MFCCs + Energy + delta + acceleration

HMM Architecture

The setting of the FB HMM is the same of the MB baseline system described in Section 3.2.1. A set of full-band models with different number of states (16, 18, 20, 22, 24 states) will be trained while the number of mixtures for each state of the FB models is kept as the same as the MB system. Although the state space cannot be expanded using method like HMM composition, we would like to see if increasing the number of states of the FB HMM directly can help improving its performance or not.

Training and Testing Environment

The full-band HMMs was trained on the clean-condition training dataset and multi-condition training dataset. The models are evaluated on the testing set which contains the four noises (car, subway, babble, exhibition hall). The word insertion penalty was individually tuned on the full-band system.

However, experiments show that the best result can be obtained without tuning any word insertion penalty, which is different from the MB system.

3.3.3 Full-Band Baseline Result: Clean-Condition Training

Table 3.7: Average word accuracy of the full-band baseline system under **clean-condition training**, varying the number of states per HMM

Noise Type	Number of States Per HMM				
	16	**18**	**20**	**22**	**24**
Car	56.76	59.03	58.71	59.77	57.79
Subway	58.69	66.93	65.13	59.93	56.23
Babble	51.35	59.43	59.64	58.71	57.04
Exhibition	60.53	68.12	65.51	64.85	62.27
Average	56.83	63.38	62.25	60.82	58.33

From the full-band baseline results tabulated in Table 3.7, it can observe that :

- Increasing the number of states of the full-band system from 16 states to 18 states can help to improve the recognition accuracy of the system. The average word error rate is reduced by about 3 - 8 % absolutely. The best result is obtained for setting the full-band system with 18 states per digit HMM.

- Further increasing the number of states of full-band system will degrade the performance of the system.

3.3.4 Full-Band Baseline Result: Multi-Condition Training

The results in clean-condition training suggested the best FB system can be obtained by the 18-state HMM. In this experiment, only the results of the 16-state and 18-state HMM will be given.

Table 3.8: Average word accuracy of the full-band system under **multi-condition training**

Noise Type	FB 16S	FB 18S
Car	89.73	90.56
Subway	90.14	91.29
Babble	88.83	89.21
Exhibition	90.18	90.88
Average	89.72	90.49

The Table 3.8 reflects that the 18-state full-band HMM can yield the best recognition result and so is chosen as the FB baseline.

3.3.5 Summary

For both the two training conditions, the FB systems perform similarly. Increase the number of states of the full-band system directly from 16 states to 18 states can help to reduce the word error rate of the system. Since the Aurora2 test data set is a noisy dataset, increasing the number of HMM states can help to better model the noise inside the speech data.

Table 3.9: Average word accuracy of the multi-band baseline system and the full-band baseline system under **clean-condition training** (**MB_w0.5_04** = MB system with SB weights [**0.5:0.5**] and asynchrony limit of four states, **FB_16S** = 16-state FB system, **FB_18S** = 18-state FB system)

Noise Type	**MB_w0.5_a4**	**FB_16S**	**FB_18S**
Car	54.89	56.76	59.03
Subway	59.89	58.69	66.93
Babble	55.93	51.35	59.43
Exhibition	53.95	60.53	68.12
Average	56.17	56.83	63.38

3.4 Overall Summary

Tables 3.9 - 3.10 summarize the baseline results for the MB system and FB systems under clean-condition training and multi-condition training, respectively. Since the Aurora2 testing data also consists of clean data [1], the performance of the MB system and the FB system under clean environment can also be compared. The recognition results of the MB system and FB system on four **clean** datasets under clean-condition training and multi-condition training are shown in Tables 3.11 - 3.12, respectively.

[1]The **clean** dataset refers to the testing data without adding the noise. Since there is a clean dataset for each noise, the 4004 utterances can be combined as a whole because the dataset does not include any noise.

Table 3.10: Average word accuracy of the multi-band baseline system and the full-band baseline system **multi-condition training**

Noise Type	**MB_w0.5_a4**	**FB_16S**	**FB_18S**
Car	88.75	89.73	90.56
Subway	88.00	90.14	91.29
Babble	86.51	88.83	89.21
Exhibition	85.80	90.18	90.88
Average	87.27	89.72	90.49

Table 3.11: The recognition result for the multi-band baseline system and the full-band baseline system on four **clean** datasets under **clean-condition training**

Noise Type	**FB_16S**	**FB_18S**	**MB_w0.5_a4**
Car	99.02	99.16	99.28
Subway	98.93	99.2	99.26
Babble	99.09	99.09	98.97
Exhibition	99.14	99.38	99.44
Average	99.05	99.21	99.24

Table 3.12: The recognition result for the multi-band baseline system and the full-band baseline system on four **clean** datasets under **multi-condition training**

Noise Type	**FB_16S**	**FB_18S**	**MB_w0.5_a4**
Car	98.45	98.69	98.69
Subway	98.68	98.86	98.99
Babble	98.76	98.91	98.79
Exhibition	98.64	98.86	99.23
Average	98.63	98.83	98.93

Chapter 4

Multi-Band System with Optimal Band-partition

4.1 Introduction

Based on the baseline results shown in Chapter 3, the full-band system performs much better than the multi-band system under noisy environments. Researchers can adjust the sub-band weights using the minimum classification error (MCE) criterion or the maximum mutual information (MMI) criterion [16] to further improve the MB system. However, the poor results from the MB system suggest that:

- The 4 different noises (car, subway, babble, exhibition hall) in Aurora2 are not band-limited. If a noise is band-limited so that it only corrupts a particular frequency band, only the acoustic features in that sub-band is corrupted while the remaining bands remain intact. Therefore, the whole MB system is believed to have a better performance over the FB system by correctly de-emphasizing the corrupted sub-band and emphasizing the other sub-band information. However, when the noise

is no longer band-limited, all sub-band acoustic features are corrupted in much the same way that the full-band features are corrupted. In this case, we expect little, if any, performance gain can be obtained by simply emphasizing/de-emphasizing certain bands.

- The assumption of sub-band independence leads to the loss of the correlation information among the sub-bands. For an MB system, the interaction between the bands are not modeled because the frequency spectrum is split into several frequency bands. However, the experiments suggested that this information is important for speech recognition [11].

- On the other hand, we hypothesize that the band partition may not be an optimal partition for the noisy speech. Figure 4.1 shows that the noise energy distributions are different for the four real noises in Aurora2. It can be observed that the noise energy usually concentrates on the lower spectral components; and for different noises, their frequency characteristics are different.

By knowing the spectrum characteristics of noise, we hope to find out the optimal band partition so that the effect of noise may be better contained. By adjusting the sub-band weights appropriately, the MB system can obtain a better performance.

However, it is hard to predict where we can partition the frequency spectrum effectively. In this chapter, A set of MB experiments are run so as to determine the optimal band partition of the 4 different noise environments. For different sub-band partitions, an MB model is trained accordingly. Based on the equal-band partition, the sub-band partition of the MB system is shifted to the lower/higher band according to the frequency spectrum of the noise.

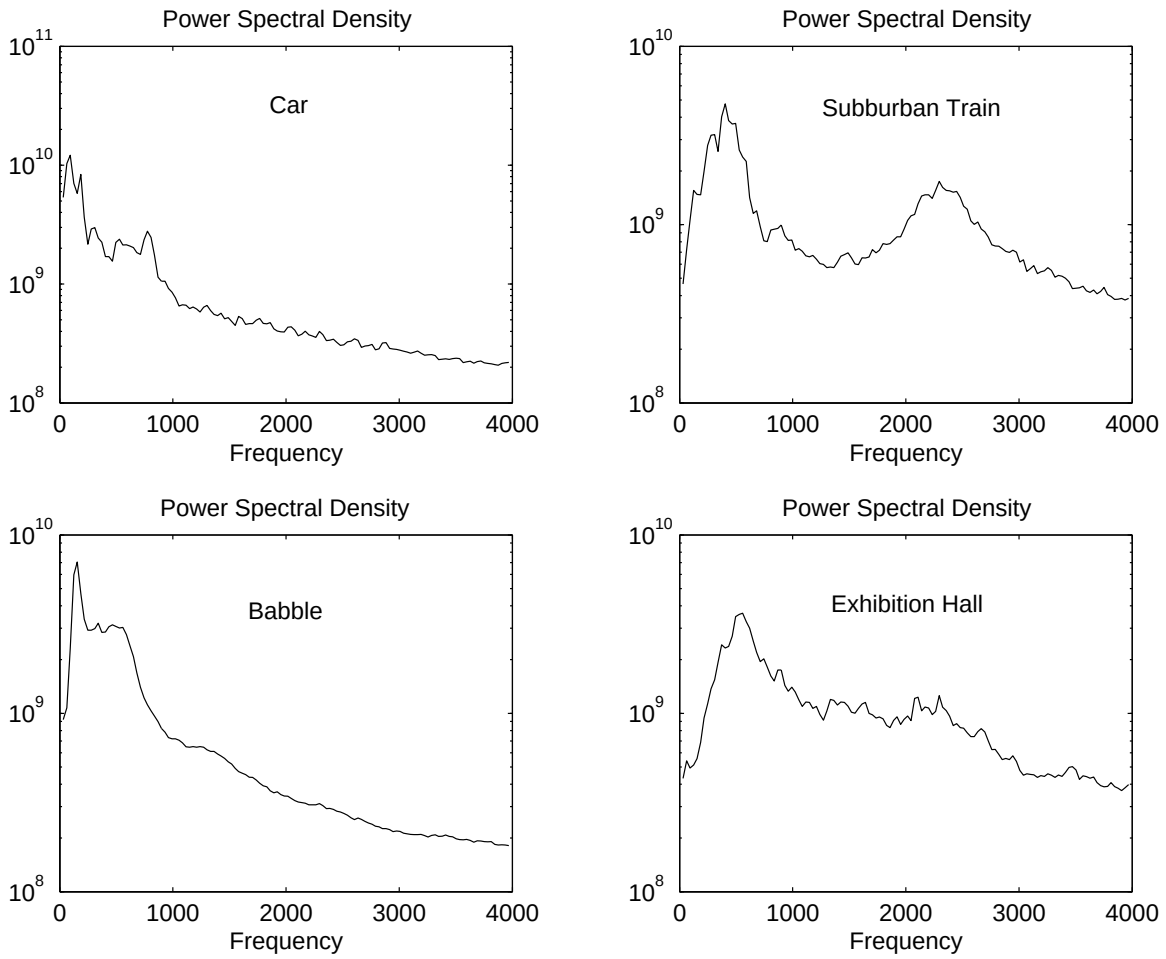

Figure 4.1: The frequency spectrum of the four noise signals

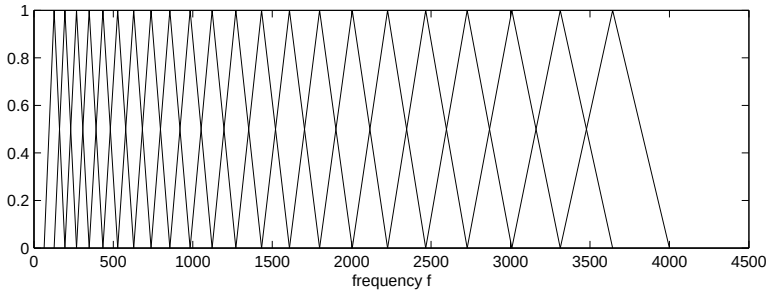

Figure 4.2: The mel-scale filterbank

The MB models are then tested to investigate the relationship between the sub-band partition and the performance of the MB system. Among different sub-band partitions, the one with the best performance is selected and its band partition will become the optimal sub-band partition of the MB system.

4.2 Experimental Setup

4.2.1 Frequency Band Partition Configurations

The band partition of the MB system cannot be shifted randomly. Under the mel-frequency scale, the whole frequency spectrum (64 Hz - 4000 Hz) is divided into 22 frequency bands. Fig. 4.2 illustrates the general form of this mel-scaled filterbank. As shown, the filters used are triangular and equally spaced along the mel-scaled defined in the following equation [25].

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{4.1}$$

According to Table 3.2, a set of band partitions have been defined. For each band partition, an MB model is trained and evaluated with the same test set. As described before, the noise energy for the noises are concentrated on the lower sub-band, which implied that the optimal band partition of

these noises should be in the lower frequency range. Table 4.1 shows the configurations of the different band partitions with different low pass cut-off frequency and high pass cut-off frequency.

Table 4.1: Configurations of the different band partitions based on the mel-scale filterbank, where **LLcut**, **LHcut** refer to low pass and high pass cut-off frequency of the first sub-band, **HLcut**, **HHcut** refer to the low pass and high pass cut-off frequency of the second sub-band

Index	**LLcut**	**LHcut**	**HLcut**	**HHcut**
L-4R+4	64	737	628	4000
L-3R+3	64	855	733	4000
L-2R+2	64	983	855	4000
L-1R+1	64	1121	983	4000
L-0R+0	64	1271	1121	4000
L+1R+1	64	1433	1271	4000

The band partition **L-0R+0** is the equal-band partition baseline system. The band partition index **L-NR+N** is the band partition that shifts **N** frequency bands left (towards the first/lower sub-band) relative to **L-0R+0**, while the band partition index **L+NR-N** is the band partition that shifts **N** frequency bands right (towards the second/higher sub-band).

4.2.2 Sub-band Feature Configurations

For the equal-band partition MB baseline system, the feature dimensions for the two sub-bands are set as equal because both of them contain the same number of frequency bands. However, after shifting the cut-off frequency of the sub-bands, the number of the frequency bands is no longer the same

for different sub-bands. By assuming each of the filterbank contains same amount of information, the feature dimension for each band partition is set according to the number of frequency bands that the sub-band belongs. Table 4.2 shows the feature dimensions of these MB systems.

Table 4.2: Configurations of feature dimensions of different band partitions (**# chans** = number of frequency bands; **# features** = dimension of the acoustic features for the sub-band model; **Total # of features** = total number of acoustic features for the two sub-band models)

Index	**Multi-Band System**				**Total # of features**
	1st sub-band		**2nd sub-band**		
	# chans	**# features**	**# chans**	**# features**	
L-4R+4	7	4 MFCCs + SB Energy + Δ + $\Delta\Delta$	15	8 MFCCs + SB Energy + Δ + $\Delta\Delta$	42
L-3R+3	8	4 MFCCs + SB Energy + Δ + $\Delta\Delta$	14	8 MFCCs + SB Energy + Δ + $\Delta\Delta$	42
L-2R+2	9	5 MFCCs + SB Energy + Δ + $\Delta\Delta$	13	7 MFCCs + SB Energy + Δ + $\Delta\Delta$	42
L-1R+1	10	5 MFCCs + SB Energy + Δ + $\Delta\Delta$	12	7 MFCCs + SB Energy + Δ + $\Delta\Delta$	42
L-0R+0	11	6 MFCCs + SB Energy + Δ + $\Delta\Delta$	11	6 MFCCs + SB Energy + Δ + $\Delta\Delta$	42
L+1R-1	12	7 MFCCs + SB Energy + Δ + $\Delta\Delta$	10	5 MFCCs + SB Energy + Δ + $\Delta\Delta$	42

4.2.3 Multi-Band Architecture and Recombination

The HMM architecture of the MB system is the same as the equal-band partition MB. The SB models are then converted into the MB model under the HMM composition framework that described in Section 3.2.1. The sub-band weights are set to 0.5 to assume equal contribution for the sub-bands.

4.2.4 Training and Testing Environment

The SB HMMs are trained with the clean-condition training dataset and evaluated on the same testing set. The word insertion penalty was individually tuned on some of the testing set for the MB system in order to obtain a better result.

4.3 Experimental Results

4.3.1 MB HMM with Optimal Band Partition vs MB HMM with Equal-Band Partition

Optimal Band Partition – For Each Noise

The recognition results for the MB systems by using different band partition settings are shown in Figures 4.3 - 4.6. Each of the figure shows the performance of the MB system under a particular noisy environment, which allow state asynchrony from zero to four. The experimental results reflect that the effect of shifting the sub-band partition is different for the four noises.

Table 4.3 summarizes the performance of MB system before/after shifting the sub-band partition to the optimal point. The **configuration** in the table refers to the configuration of the MB system with the best result. For the

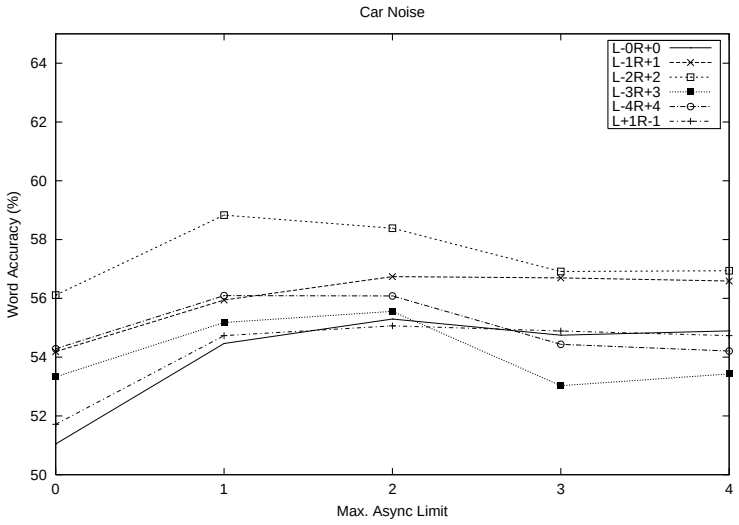

Figure 4.3: The recognition results for the multi-band system on car noise under **clean-condition training**, with maximum asynchrony limit up to four states

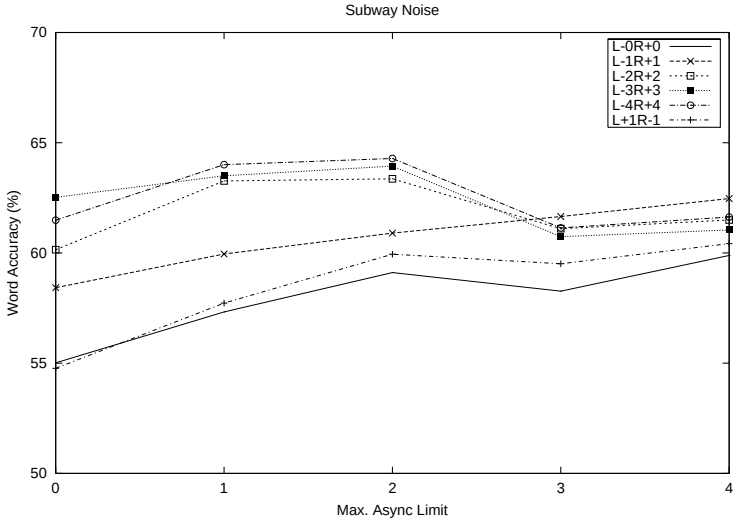

Figure 4.4: The recognition results for the multi-band system on subway noise **clean-condition training**, with maximum asynchrony limit up to four states

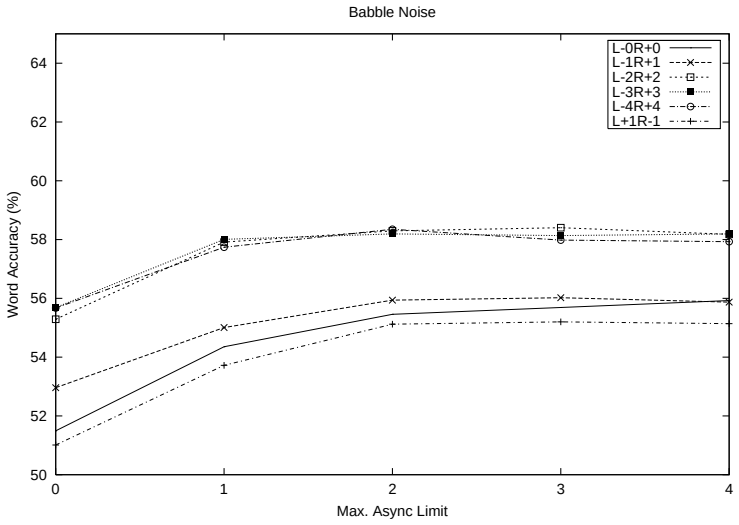

Figure 4.5: The recognition results for the multi-band system on babble noise **clean-condition training**, with maximum asynchrony limit up to four states

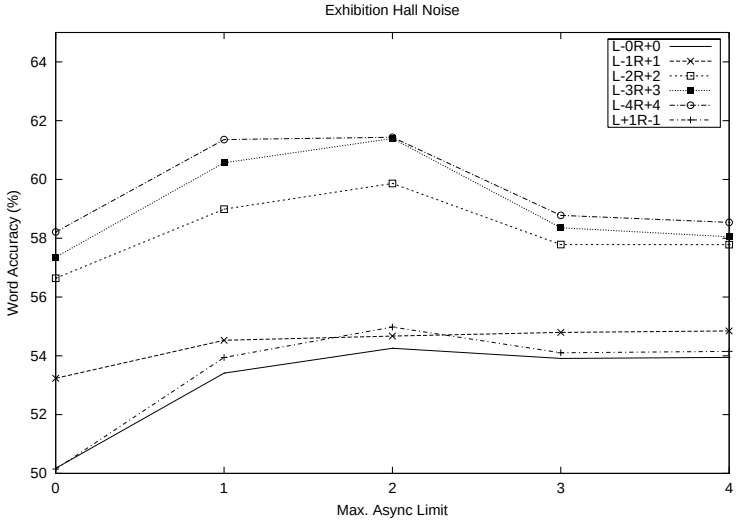

Figure 4.6: The recognition results for the multi-band system on exhibition hall noise **clean-condition training**, with maximum asynchrony limit up to four states

case of car noise, the best result is obtained by shifting the sub-band partition to **L-2R+2** with limiting the asynchrony to 1 state.

From the Table 4.3, it shows the performance gain for shifting the sub-band partitions differently for different noises. The performance gain varies from 2.5% to 7.2% absolutely, with an average gain of 4.4%.

Table 4.3: Average word accuracy between the optimal band partition multi-band system and the equal-band partition multi-band system

	Optimal Band Partition		**Equal-Band Partition**	
Noise	**configuration**	**word acc.**	**configuration**	**word acc.**
Car	L-2R+2, a1	58.83	L-0R+0, a2	55.3
Subway	L-4R+4, a2	64.28	L-0R+0, a4	59.89
Babble	L-2R+2, a3	58.4	L-0R+0, a4	55.93
Exhibition	L-4R+4, a2	61.44	L-0R+0, a2	54.26

Optimal Band Partition – For All Noise

The average recognition results of the four noises are plotted in Figure 4.7. For each noise, the average word accuracy is calculated on five of the seven subsets, which were the range from **20 dB** dataset to **0 dB** dataset. The **clean** and **-5 dB** subsets in the whole testing set are excluded.

From Figure 4.7, several points can be observed:

1. The "optimal" band partition should be somewhere around **L-2R+2**, **L-3R+3**, **L-4R+4**. The best result is obtained from the MB system using the band partition **L-4R+4** and limiting maximum asynchrony

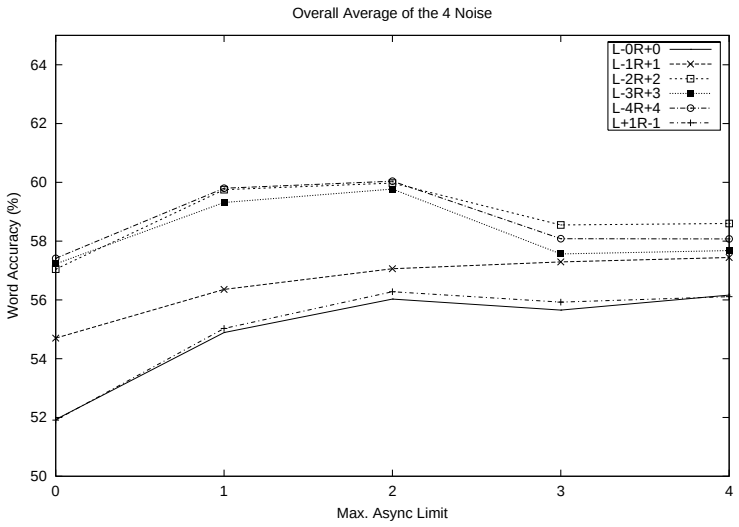

Figure 4.7: Average recognition results for the MB system on the four noises under **clean-condition training**, with maximum asynchrony limit up to four states

of two states. However, the band partition **L-2R+2** gives the best result on the overall.

2. The MB baseline system, with the band partition **L-0R+0** gives the worst recognition result. Shifting the sub-band partition slightly to the left or right (**L-1R+1**, **L+1R-1**) only gives slightly better performance. The relative performance gain for **L-1R+1** from **L-0R+0** is much larger than that for **L+1R-1** from **L-0R+0** suggests that the band partition should be shifted towards the lower band.

3. When the band partition is close to the "optimal" point (**L-2R+2**, **L-3R+3**, **L-4R+4**), the MB systems' performance are similar. The best recognition result is obtained by limiting the maximum asynchrony L to two states. Further relaxing asynchrony results in degraded performance. However, for the remaining band partition (**L-1R+1**, **L-**

0R+0, **L+1R-1**), increasing L will help to improve the performance of the system.

4. Although the asynchrony effect was different by using different band partitions and no definite conclusion can be drawn on whether the asynchrony limit should be stopped, it can be observed that further relaxing asynchrony to more states can still maintains the performance to be as good as the synchronous system ($L = 0$).

Optimal Band Partition Multi-Band System

Among different settings, the band partition **L-2R+2** is chosen to be the optimal band partition for the four noises. The best recognition result is given by limiting the maximum asynchrony limit to two states, it is chosen to represent the best result our MB system can give.

4.3.2 MB HMM with Optimal Band Partition vs FB HMM

The experimental results show that there is improvement for shifting the sub-bands partition to an optimal point. The comparison between the MB system and the FB system is shown in Table 4.4.

In the table, **MB_w0.5_a2** refers to the MB system with its sub-band weight of the lower sub-band set to 0.5 (the sub-band weight of the higher sub-band will be $1 - 0.5 = 0.5$), and the maximum asynchrony limit to two states.

The table shows that the 16-state FB system give the worst results. Although the MB system with the optimal band partition outperforms the 16-state FB system, its performance is not as good as that of the 18-state

Table 4.4: Average word accuracy of the multi-band system with optimal band partition **L-2R+2** setting and the 16-state, 18-state full-band system

Noise	**FB_16S**	**MB_w0.5_a2**	**FB_18S**
Car	56.76	58.38	59.03
Subway	58.69	63.36	66.93
Babble	51.35	58.3	59.43
Exhibition	60.53	59.86	68.12
Average	56.83	59.98	63.38

FB system. Among the four noises, the MB system gives the best result relatively for the the car noise testing set. The difference of the average word accuracy between the two systems is less than 1% absolutely for this case. The case of exhibition hall noise gives the worst result because the FB system performs much better than the MB system.

4.4 Discussions and Conclusions

4.4.1 Discussions

Among the results, several points are worth discussing:

- When the sub-band partition of the MB system is set as **L-0R+0**, its performance is poor. Referring to the frequency spectrum of the four noises shown in Figure 4.1, it seems that the **L-0R+0** partition may not be an optimal partition. With this band partition, the noise and speech cannot be "separated" effectively. It is hypothesized that the optimal band partition should be somewhere inside the lower band. By shifting the sub-band partitions to the lower band, the performance of

the MB system improved accordingly until the sub-band partition is an optimal point.

The idea of shifting the sub-band partition is similar to adjusting the sub-band weights of the MB system. The two approaches re-adjust the contribution of the sub-bands, with the hypothesis that the equal contribution of the sub-bands is inappropriate under noisy environments. The main difference is that for the case of tuning the sub-band weights, the sub-band information are fixed. The system determines which sub-band is more informative and reliable, and emphasized that particular sub-band. However, for the case of shifting sub-band partition, the sub-band information can be changed. After shifting the sub-band partition, the new frequency partition is believed to have "separated" the noise and speech information in a more effective manner. The new MB system can give a satisfactory performance even when the sub-band weights are set to be equal.

- For the MB system with its sub-band partition far away from the optimal point, it is interesting to observe that there is still rooms for improvement for further relaxing asynchrony. However, it is not the case for the band partition that is closed to the optimal point, further relaxing asynchrony will degrade the performance. This may due to the performance gain from the asynchrony effect and that the gain from shifting band partition are not addictive. For a "poor" MB system (with bad sub-band partition), relaxing asynchrony can lead to a large improvement for the performance, however, the performance gain will be smaller for the "good" MB system because the system has already reached a satisfactory result.

4.4.2 Conclusions

- Under noisy environments, the equal-band partition MB system performs poorly which proposes that the band partition **L-0R+0** is not an optimal band partition. The performance gain of the MB system using the band partition **L-2R+2** suggests that the optimal partition should be somewhere near this vicinity.

- Shifting the sub-band partition of the MB system is similar to the idea of adjusting the sub-band weights. However, the former approach can result in a significant gain even though the sub-band weights are not adjusted. It is thus an effective approach for improving the MB system.

- It is observed that the performance gain of the MB system from shifting sub-band partition to an optimal point and relaxing the asynchrony between the sub-band transitions may not be addictive.

- The performance of the MB system with optimal band partition can perform better than the 16-state FB system. However, it is still not as good as that of the 18-state full-band system. It is suggested that the MB system is not good enough to deal with real noise because (1) it cannot make use of the sub-band independence property to separate speech and noise information effectively; and (2) it loses the correlation information among the sub-bands.

Chapter 5

Frequency-Stream-Tying Hidden Markov Model

5.1 Motivation

Besides the non-optimal band partition problem, another weakness of the multi-band (MB) system we used is the training process of the sub-band (SB) models. For a conventional MB system, researchers will split the whole frequency into sub-bands and train a model for each band. They will then focus on the recombination of the sub-band information, either by linear or non-linear recombination.

However, for that kind of training, the SB models can only be optimized independently because the alignments for updating the model parameters during the training process are independently found and are different for the sub-bands. The recombination of the SB models to the MB model will give rise to the question: Is the MB model being optimized? It is questionable that although the SB models are being optimized through the training process, it may not result in the optimization of the MB model, which is a combination

of the SB models.

It would be interesting if we can develop a new mechanism/algorithm so that the new system can still keep the properties of the MB system, with the addition of jointly optimizing the SB models. The problem can be solved by the use of the frequency-stream-tying (FST) HMM. By the use of this new proposed system, the SB model parameters can be jointly optimized through the training process.

The idea is first to design an HMM topology that exactly equals to the composite HMM generated by the SB models. The FSTHMM has multiple streams of acoustic features, and each stream consists of acoustic features extracted from one sub-band. These streams are tied across states as in an MB HMM created by HMM composition to allow asynchronous transitions of different sub-band acoustic features. Through the tying of the stream elements, the FSTHMM can be interpreted as a MB HMM. Since the FST model is also a kind of HMM, the whole model, which consists of sub-band parameters, can be jointly optimized through the conventional training algorithm such as Baum-Welch algorithm, at the same time.

In summary, the following characteristics of the frequency-stream-tying HMM should be noted:

- It will separate acoustic sub-bands using the same approach as the MB system. The optimal band partition can be applied to the new system directly.

- It can allow asynchrony for sub-band transitions, by tying on the streams of the FSTHMM, to model the speech signal more accurately.

- The new HMM which consists of the SB parameters can be jointly optimized through the training process.

- The new model is also a type of HMM, which means it can be applied directly on HMM training and decoding algorithms, such as Baum-Welch algorithm and Viterbi algorithm, without any modifications on the algorithms.

- Similar to the MB system, the stream weights (sub-band weights) can be optimized afterward to further enhance system's performance.

5.2 Frequency-Stream-Tying HMM

5.2.1 Related Works

FSTHMM divides the states into independent streams representing the sub-band information. Logan *et al.* [26] has proposed an extended HMM namely the factorial hidden Markov model (FHMM). By dividing the states into layers, the FHMM can model several stochastic random processes with loosely coupled dynamics. The FHMM can be represented as a dynamic belief network [27] composed of several "layers" and each layer has independent dynamics. The layers have similar dynamics to a basic HMM but the probability of an observation at each time depends upon the current state in all of the layers. The new system seems to allow a more natural fit to speech.

Glotin *et at.* [28] has proposed a multi-stream HMM which combines the audio and visual modalities (streams). They take a decision fusion approach for the audio-visual information where the single audio- or visual-HMM classifiers are combined to recognize the audio-visual speech. There are two different techniques for the combination of the stream information. The first technique estimates the audio stream reliability to adjust the weights for each modality, while the second one use the Discriminative Model Combination

(DMC) approach.

When the test data and train data used to train the automatic speech recognition (ASR) system come from different acoustic environment, the performance of the ASR system degrades. To address the problem of this mismatch environment, Shire [29] have proposed a multi-stream HMM in which the constituent streams are trained in separate acoustic environments. The probabilities estimated from the classifiers will be combined in different ways. It is reported that when the streams are trained separately with clean data and heavily reverberated data, the combined system can improve the ASR performance with unseen reverberated test data.

5.2.2 Topology of the Frequency-Stream-Tying HMM

Previous researchers separate the feature vectors into different streams for the cases of building the tie-mixture system or using the vector quantization. With this approach, the whole feature vectors is divided into several streams by separating the energy, deltas components, etc. For our proposed frequency-stream tying system, it contains multiple streams of acoustic features. However, the stream corresponds to the acoustic features extracted from one particular sub-band.

As shown in Figure 5.1 and Figure 5.2, the topology of the FSTHMM is exactly the same as that of the MB system. The main difference between the two systems is that the MB HMM is generated by the sub-band HMMs using the HMM composition algorithm, while the FSTHMM is directly trained through the training process. For instance, by applying the HMM composition, the 3-state MB HMM with maximum asynchrony limit of 2 states ($N_{mb} = 3, A = 2$) will be generated as a composite HMM with number of states equals to 9. To train an FSTHMM with the same configurations as the

MB HMM ($N_{fst} = 3, A = 2$), we first design a 9-state HMM with the same topology as the MB HMM. For each state of the FSTHMM, there are two streams, with stream 1 and stream 2 representing the lower sub-band and the higher sub-band, respectively. The FSTHMM will not be a left-to-right HMM, but with the state transitions shown in the Figure 5.2. The remaining procedure is to tie the streams across the states.

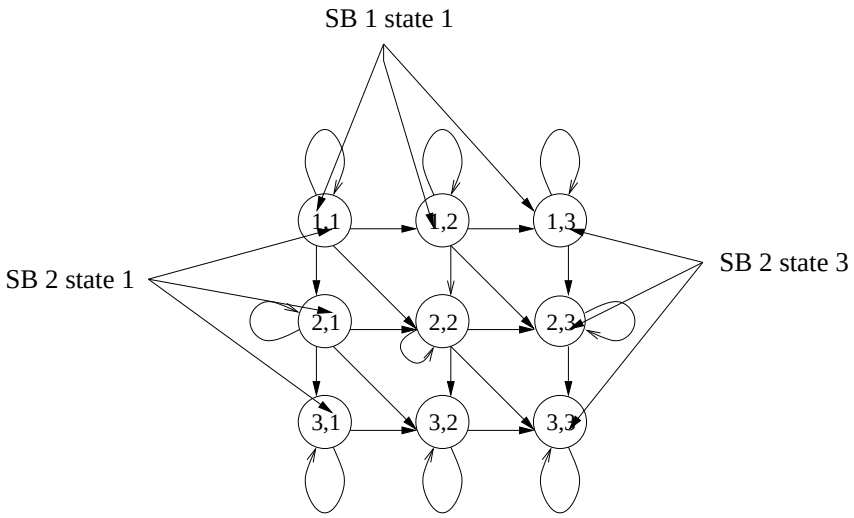

Figure 5.1: A multi-band HMM, with number of states $N_{mb} = 3$, asynchrony limit $A = 2$

As shown in Figure 5.2, the state (1,1), state (2,2) and state (3,3) belong to the case of the FSTHMM with fully synchrony. Stream 1 for state (1,1), state (1,2) and state (1,3) of the HMM corresponds to sub-band 1 of state 1 of the FSTHMM, and so are tied together to share the same distributions. It means that the lower sub-band of the FSTHMM is in state 1, no matter what state the higher sub-band belongs to. Stream 2 for state (1,1), state (2,1), state (3,1) of the HMM corresponds to sub-band 2 of state 1 of the FSTHMM, with the meaning that the higher sub-band of the FSTHMM is

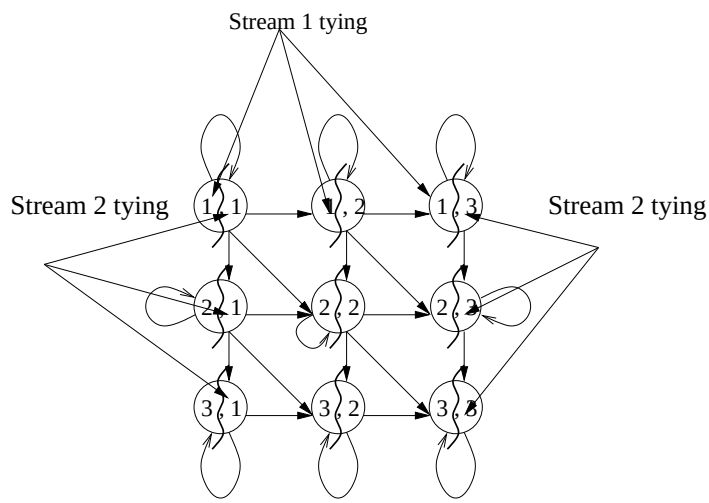

Figure 5.2: A frequency-stream-tying HMM, with number of states $N_{fst} = 3$, asynchrony limit $A = 2$

in state 1, no matter what state the lower-sub-band belongs to. The detail relationships are shown below.

$$\begin{cases} s^1_{(1,1)}, s^1_{(1,2)}, s^1_{(1,3)} & = & S^1_{fst,1} \\ s^1_{(2,1)}, s^1_{(2,2)}, s^1_{(2,3)} & = & S^1_{fst,2} \\ s^1_{(3,1)}, s^1_{(3,2)}, s^1_{(3,3)} & = & S^1_{fst,3} \\ s^2_{(1,1)}, s^2_{(2,1)}, s^2_{(3,1)} & = & S^2_{fst,1} \\ s^2_{(1,2)}, s^2_{(2,2)}, s^2_{(3,2)} & = & S^2_{fst,2} \\ s^2_{(1,3)}, s^2_{(2,3)}, s^2_{(3,3)} & = & S^2_{fst,3} \end{cases} \text{where,}$$

$$s^a_{(i,j)} = \text{stream a of the state (i,j) for the 9-state HMM,}$$

$$S^a_{fst,k} = \text{sub-band a of the state k for the 3-state FSTHMM with,}$$

$$1 \leq i, j \leq 3,\ 1 \leq k \leq 3,\ a = 1, 2$$

With knowing the relationships between the streams for each state, all the streams that belong to the same sub-band and same states will be tied together to pass the training process. With this stream tying mechanism, a

frequency-stream-tying HMM with $N_{fst} = 3$ and $A = 2$ can be trained with the sub-band parameters jointly optimized through the training process.

5.2.3 Training Procedures of the Frequency-Stream-Tying HMM

The training procedures of the frequency-stream-tying HMM with 2 streams are shown in the following steps:

1. Given the number of states N_{fst} and maximum asynchrony limit A of the FSTHMM, the total number of states N the HMM needed is computed.

2. Define an N states HMM prototype, with 2 streams per each state. The stream (sub-band) weights are set to be equal for simplicity. The transitions between the states depend on the maximum asynchrony limit A, as shown in Figure 5.2.

3. Stream tying is applied to the streams of particular states. The number of parameters is nearly the same as the MB model and the full-band model after the tying.

4. Before the training, a flat-start model will be determined such that the Gaussian parameters for all the models are the same.

5. Apply the training by the Baum-Welch Algorithm on the FSTHMM until the model parameters converge.

6. Increase the number of mixtures per stream by 1, and train the model until it converges.

7. Split the mixtures per stream to the fixed value M (for our case, $M = 3$), and train the model until it converges.

5.2.4 Decoding of the Frequency-Stream-Tying HMM

Through the training, the parameter sets Θ of the FSTHMM, where $\Theta = (A, B, \pi)$ corresponds to the state transition probabilities, probability distributions and the initial state probabilities, respectively, will be trained. The FSTHMM is used to perform recognition by obtaining the single best state sequence $\overline{Q} = (\overline{q}_1\ \overline{q}_2\ ...\overline{q}_T)$. This maximizes $P(Q, O|\Theta)$ given the observation vectors O and HMM parameters Θ of all possible state sequences Q.

The decoding process will use the Viterbi algorithm to obtain the single best state sequence. The definition of $\delta_t(j)$, which is the highest score along a single path which ends at state j at time t will be given through the decoding process. $\delta_t(j)$ can be recursively computed from time 0 to t by:

$$\delta_t(j) = \max_{\overline{q}_1\ \overline{q}_2\ ...\overline{q}_T} P[\overline{q}_1\ \overline{q}_2\ ...\overline{q}_{t-1}, \overline{q}_t = j, o_1^t|\Theta] \tag{5.1}$$

$$= \max_j [\delta_{t-1}(i)a_{ij}]b_j(o_t) \tag{5.2}$$

The observation probability, $b_j(o_t)$, is the probability that an observation feature vector o_t at time t belongs to state j and it can be expressed as:

$$b_j(o_t) = P(o_t|j, \Theta) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(o_t; \mu_m, \Sigma_m) \tag{5.3}$$

where M is the number of mixture components in state j, c_{jm} is the weight of the m-th component and $\mathcal{N}(\cdot; \mu_m, \Sigma_m)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ.

For the frequency-stream-tying HMM, the $b_{j,l}(o_t)$ can be expressed in a more general form:

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm}\mathcal{N}(o_{st}; \mu_{sm}, \Sigma_{sm}) \right]^{\gamma_s} \tag{5.4}$$

where M_s is the number of mixture components in state j for stream s, c_{jsm} is the weight of the m'th component.

5.3 Experiment I: Frequency-Stream-Tying HMM under Clean-Condition Training

In the previous chapter, we have trained a set of MB models with different sub-band partitions. The results show that the band partition **L-2R+2** gives the best result. However, the results show that the MB system, even with optimal band partition, cannot obtains a good performance compared to the full-band system. In this chapter, the FSTHMM is trained and evaluated using the same dataset. It is hypothesized that the new system, the frequency-stream-tying HMM, should perform better than the MB system.

5.3.1 Experimental Setup

A set of FSTHMMs with maximum asynchrony limit up to 3 states will be trained using the band partition **L-2R+2**. The FSTHMM is trained following the steps listed in Chapter 5.2.3. Like the MB models, there will be 3 mixtures for each of the streams in each of the states of the digit model and 6 mixtures for each of the pause models, which is described in Chapter 3.2.1.

5.3.2 Experimental Result

Table 5.1 shows the average word accuracy of the frequency-stream-tying system. The maximum asynchrony limit will be ranging from zero to three states.

It is observed that the asynchrony effect for the frequency-stream-tying

Table 5.1: The average word accuracy of the frequency-stream-tying system with maximum asynchrony limit ranging from zero to three states under **clean-condition training**

Noise Type	Maximum Asynchrony limits			
	a0	**a1**	**a2**	**a3**
Car	58.28	58.78	58.11	59.03
Subway	63.79	64.91	65.01	63.24
Babble	57.64	57.59	57.06	58.65
Exhibition	60.06	61.17	59.57	60.39
Average	59.94	60.61	59.94	60.33

system is much smaller than that of the multi-band system. The best result can be obtained by limiting the asynchrony of the FST system to 1 state.

Table 5.2 compares the average recognition result of the frequency-stream-tying system with the multi-band system and the full-band system. The average recognition results for each noise is plotted in Figures 5.3 - 5.6.

From the results, it can be observed that:

1. The average performance gain for the FST system over the MB system is about 0.6% absolutely. The gain seems to be small. However, the performance gain varies differently across different noises. For the babble noise, the MB system performs better than the FST system when the maximum asynchrony limit is fixed (**a1**). For other noises, the gain for the FST system over the MB system is about 1% - 1.5%. The statistical significance tests in Appendix D show that the FST system is statistically significant better than the MB system.

2. The FST system outperforms the 16-state FB system. However, when

Table 5.2: Comparison of the average word accuracy between the FST system, MB system and the FB system on different noises under **clean-condition training** (**MB_w0.5_a2** = MB system with sub-band weights [**0.5:0.5**] and asynchrony limit of two states, **FST_w0.5_a1** = FST system with sub-band weights [**0.5:0.5**] and asynchrony limit of one state, **FST Best** = best result the FST system can achieve by allowing different asynchrony limits on different noises, **FB_16S** = 16-state FB system, **FB_18S** = 18-state FB system)

Noise Type	**FB_16S**	**MB_w0.5_a2**	**FST_w0.5_a1**	**FST Best**	**FB_18S**
Car	56.76	58.38	58.78	59.03	59.03
Subway	58.69	63.36	64.91	65.01	66.93
Babble	51.35	58.30	57.59	58.65	59.43
Exhibition	60.53	59.86	61.17	61.17	68.12
Average	56.83	59.98	60.61	60.96	63.38

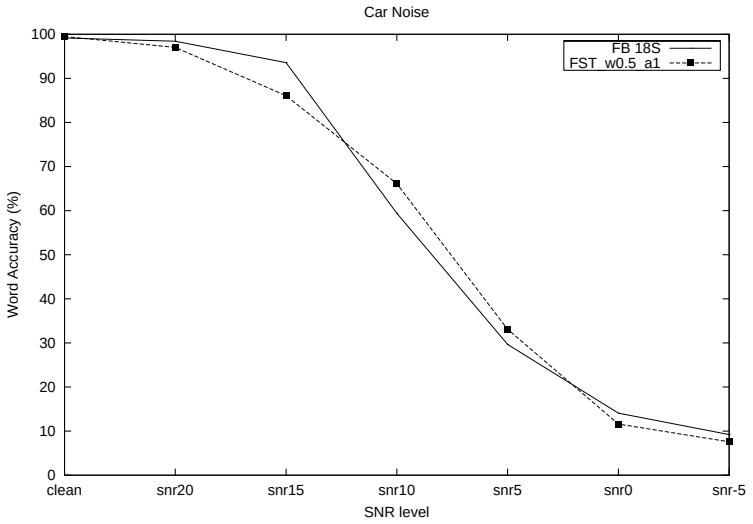

Figure 5.3: Comparison of the recognition results between the FST system and the FB system on car noise under **clean-condition training** (**FST_w0.5_a1** = FST system with sub-band weights [**0.5:0.5**] and asynchrony limit of one state, **FB 18S** = 18-state FB system)

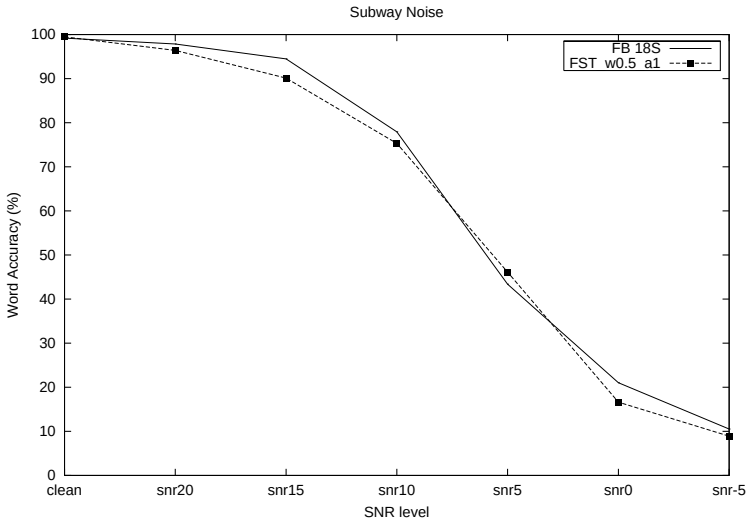

Figure 5.4: Comparison of the recognition results between the FST system and the FB system on subway noise under **clean-condition training**

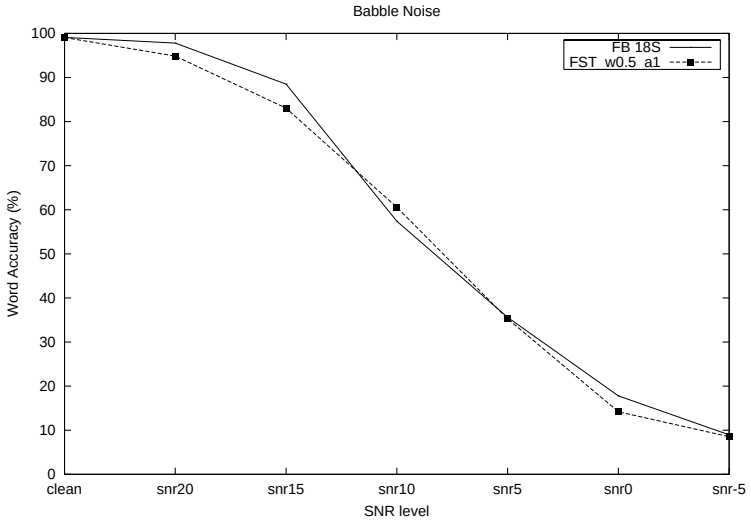

Figure 5.5: Comparison of the recognition results between the FST system and the FB system on babble noise under **clean-condition training**

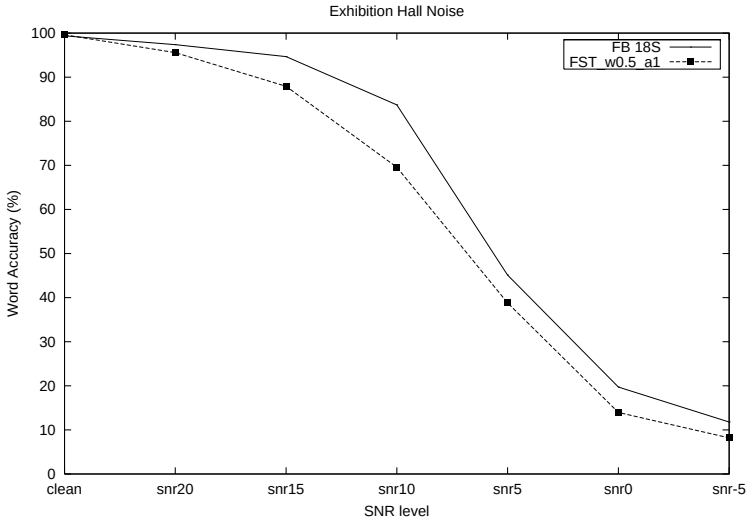

Figure 5.6: Comparison of the recognition results between the FST system and the FB system on exhibition hall noise under **clean-condition training**

comparing the recognition results for the FST system and the 18-state FB system, the FB system can give better results on the overall. The difference between the two systems for the car noise is very small. However, for the exhibition hall noise, the FB system still performs much better than the FST system.

3. Refer to Figures 5.3 - 5.6, it is observed that for noisy datasets (exclude **clean**), the FB system always outperforms the FST system for high SNR datasets (**20 dB**, **15 dB**) and low SNR datasets (**0 dB**, **-5 dB**). For the **10 dB** and **5 dB** dataset, the FST system can perform better than the FB system for some noise environments.

4. For **clean** dataset, the FST system always achieves better performance than the MB system, which outperforms the FB system. Further discussions will be given in next section.

5.4 Experiment II: Frequency-Stream-Tying HMM under Multi-Condition Training

5.4.1 Motivation : Matched Condition Between the Training and Testing Environments

Experiments in the previous section have showed that the recognition results of the FB system is better than the FST system on average. However, it is shown that the FST system always performs better than the MB system and the FB system on the **clean** dataset. Table 5.3 summarizes the recognition results for the three systems on the **clean** dataset.

It can be found that when testing on the **clean** dataset, the FST system

Table 5.3: Comparison of the average word accuracy between the FST system, MB system and the FB system on **clean** dataset under **clean-condition training**

Noise Type	**FB_16S**	**FB_18S**	**MB_w0.5_a2**	**FST_w0.5_a1**
Car	99.02	99.16	99.31	99.49
Subway	98.93	99.2	99.36	99.48
Babble	99.09	99.09	98.91	99.06
Exhibition	99.14	99.38	99.54	99.63
Average	99.05	99.21	99.28	99.42

performs the best, followed by the MB system and the FB system give the worst result. Although the performance of the FB system can be improved by increasing the number of HMM state from 16 to 18, both the FST system and MB system outperform the FB system. In terms of relative word error rate, the MB system has decreased 8.8% compared to the 18-state FB system. For the FST system, the word error rate has decreased 26.6% relatively.

For a recognizer trained on **clean** dataset and testing on the **clean** dataset, the situation is known as match-condition. It is hypothesized that the FST system and the MB system can outperform the FB system at match-condition. In the coming experiments, the same FSTHMM are trained, however, using the multi-condition training dataset. Since the multi-condition training dataset also contains the four noises with different signal-to-noise ratio, the situation can be known as match-condition.

5.4.2 Should Optimal Band Partition still Work in Multi-Condition?

The results in Chapter 4 have shown that the performance of the FST system is better than that of the MB system. Therefore, the following experiments only concentrate on the FST system.

The frequency-stream-tying HMM is trained under the optimal band partition **L-2R+2**. Since all the previous experiments are based on the clean-condition training dataset, it is interesting to examine if the optimal band partition would still work under the multi-condition. For sanity check, the FSTHMM will also be trained under the band partition **L-0R+0** to investigate the effect of shifting the band partition for an FST system. Table 5.4 and Table 5.5 show the recognition results for the two FST systems.

Table 5.4: The average word accuracy of the frequency-stream-tying system, with band partition **L-0R+0** and maximum asynchrony limit ranging from zero to three states under **multi-condition training**

	Maximum Asynchrony limits			
Noise Type	**a0**	**a1**	**a2**	**a3**
Car	88.90	88.98	89.00	89.08
Subway	90.19	90.11	90.39	90.38
Babble	86.94	87.12	86.91	86.94
Exhibition	87.46	87.65	87.35	87.61
Average	88.38	88.46	88.41	88.50

From the recognition result, it is observed that the FST system with optimal band partition outperforms the FST system with band partition **L-0R+0**. However, the asynchrony effect for the FST system is much smaller

Table 5.5: The average word accuracy of the frequency-stream-tying system, with band partition **L-2R+2** and maximum asynchrony limit ranging from zero to three states under **multi-condition training**

	Maximum Asynchrony limits			
Noise Type	**a0**	**a1**	**a2**	**a3**
Car	89.52	90.17	90.33	90.28
Subway	90.69	90.37	90.06	90.08
Babble	87.51	87.82	87.72	87.91
Exhibition	88.34	88.55	88.41	88.87
Average	89.02	89.23	89.13	89.29

than that of the MB system. For the FST systems with different asynchrony limits, their performance are similar. The best result can be obtained by limiting the FST system with the maximum asynchrony of 3 state. However, to consider both the accuracy and the complexity of the FST model, the FST system with band partition **L-2R+2** and maximum asynchrony limit of 1 state is chosen to be the best FST system to compare with the FB system.

5.4.3 Optimization of the sub-band Weights Using Minimum Classification Error Criterion

The FST system can be further improved by optimizing the sub-band weights using the string-based Minimum Classification Error (MCE) criterion. The formulae included have been derived in Chapter 2.

MCE Training Process

After the training of the FSTHMM, the sub-band weights of the FST system are trained using the string-based MCE criterion. The same training data (multi-condition training dataset) is used as the training data for the MCE training.

To perform the MCE training on continuous speech recognition, we have to determine the competing hypotheses against the correct hypothesis. In this thesis, the one-nearest string competitor described in Section 2.1.3 will be used through the MCE training.

Experimental Results

The FST system with optimal band partition **L-2R+2** and asynchrony limit of 1 state (**FST_w0.5_a1**) is chosen for sub-band weights optimization under the MCE criterion.

The MCE parameters are set during the MCE training:

- The learning rate of the gradient descent procedure at the i-th iteration $\epsilon_i = \epsilon_0 * N * (1 - i/I)$, where $\epsilon_0 = 0.001$, $N = 8440$ (total utterances), $I = 10$ (total iterations).
- The slope of the sigmoid function $\gamma = 0.1$.
- The offset of the sigmoid function $\theta = 0$.

The sub-band weights derived from the MCE training are [**0.56:0.46**]. The FST system with the MCE-derived sub-band weights is evaluated to study the effectiveness of the optimized the sub-band weights. For sanity check, an FST system with sub-band weights [**0.46:0.56**] is also evaluated. The results are shown in Table 5.6.

Table 5.6: Comparison of the average word accuracy between the FST system, FST system with MCE-derived sub-band weights, the FB system

Noise Type	Frequency-Stream-Tying System			Full-Band System	
	w0.44_a1	**w0.5_a1**	**w0.56_a1**	**16S**	**18S**
Car	89.73	90.17	90.25	89.73	90.56
Subway	89.31	90.37	91.05	90.14	91.29
Babble	87.28	87.82	88.01	88.83	89.21
Exhibition	87.83	88.55	89.09	90.18	90.88
Average	88.54	89.23	89.60	89.72	90.49

The table shows that there is a performance gain after incorporating the MCE-derived sub-band weights in the FST system. Similar to the asynchrony effect, the performance gain is small. The poor result from using the sub-band weights [**0.44:0.56**] proves that MCE-derived sub-band weights is useful and effective to improve the performance of the FST system.

Comparing the performance of the FST system with the FB system, the performance of the FST system is better than that of 16-state FB system, and can maintain a similar or slightly degraded performance for the cases of car noise and subway noise compared to the 18-state FB system. For the babble noise and exhibition hall noise, both the 16-state and 18-state FB systems give a better performance than the FST system.

The average recognition results between the FST system with MCE-derived sub-band weights and the FB system for each noise are plotted in Figure 5.7 - 5.10. The [**dataset1:dataset2**] in the x-axis refers to the recognition result that is averaged from the dataset **dataset1** and dataset **dataset2**.

Through this kind of presentation, the column [**clean:20dB**] is the recognition result averaged by **clean** dataset and **20 dB** dataset, while the column [**clean:-5dB**] is the average recognition result for all the 7 datasets.

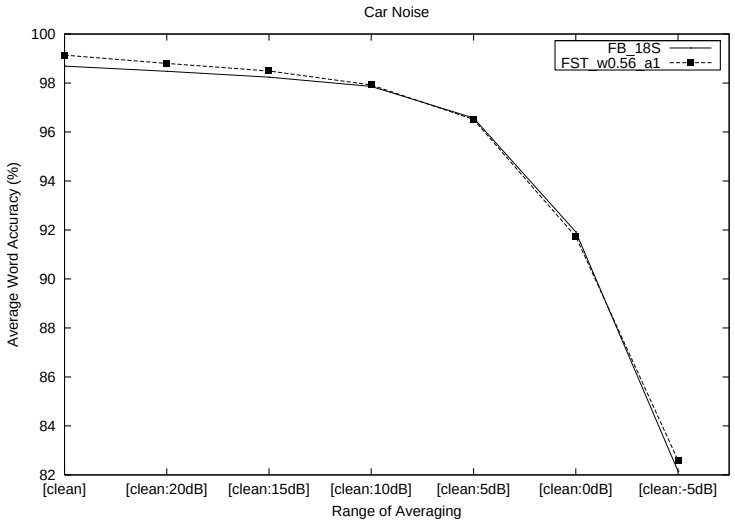

Figure 5.7: Comparison of the recognition results between the FST system with MCE-derived SB weights and the FB system on car noise under **multi-condition training** (**FST_w0.56_a1** = FST system with MCE-derived SB weights [**0.56:0.44**] and asynchrony limit of one state)

Discussions

Based on the figures, several points are worth discussing:

1. It is observed that the performance gain for the FST system after incorporating the MCE-derived SB weights is small. The MCE-derived SB weights, which shift the weight of the lower sub-band from **0.5** to **0.56**. The small change in the sub-band weights can only lead to a small improvement in performance of the FST system.

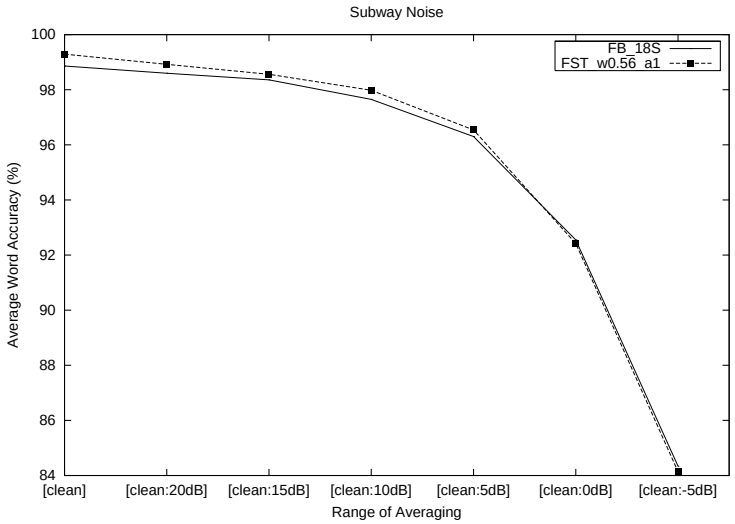

Figure 5.8: Comparison of the recognition results between the FST system with MCE-derived SB weights and the FB system on subway noise under **multi-condition training**

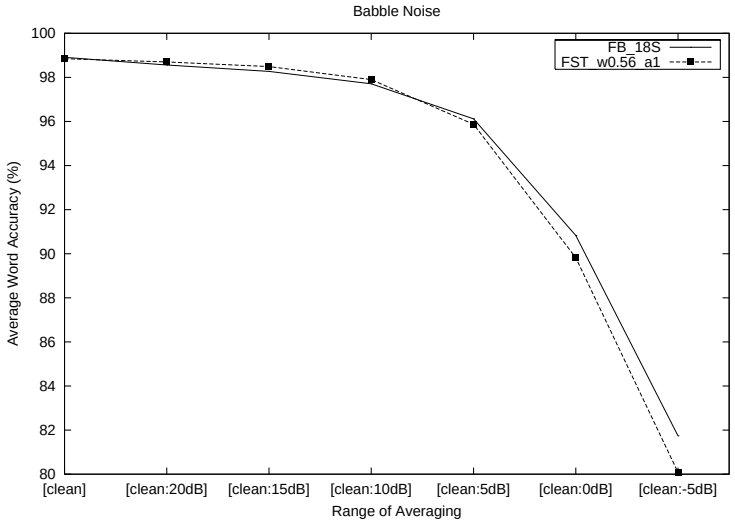

Figure 5.9: Comparison of the recognition results between the FST system with MCE-derived SB weights and the FB system on babble noise under **multi-condition training**

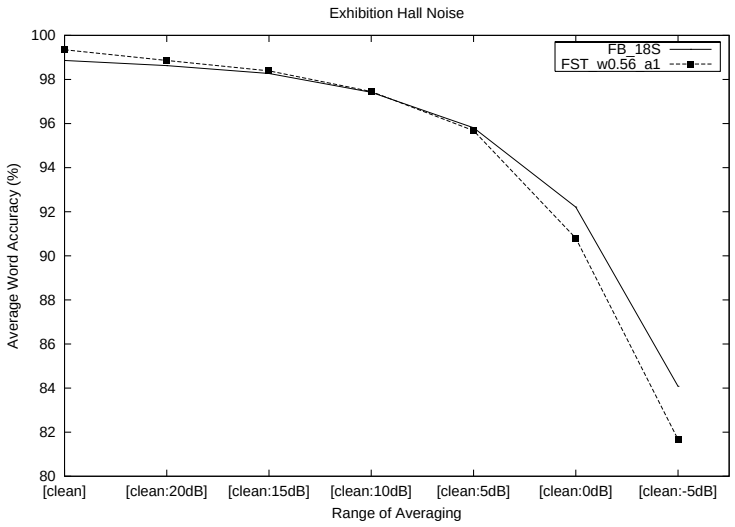

Figure 5.10: Comparison of the recognition results between the FST system with MCE-derived SB weights and the FB system on exhibition hall noise under **multi-condition training**

The SB weights for the FST system can only be shifted by a small value is mainly due to the training procedure of the FST system. For an FST system, the SB (stream) weights are initially set as equal to passed through the training process. Through the Baum-Welch training algorithm, the model parameters of the FST system are being optimized under the criterion of uniform SB weights. As a result, further optimization of the SB weights on an FST model, which has been optimized by setting the SB weights to be equal, can only lead to a small change in the SB weights.

2. Among the recognition results for the four noises, the FST system performs better for the car noise and subway noise, which can be grouped as **machine noise**. The overall recognition result of the FST system with the MCE-derived SB weights is similar to the FB system on **ma-**

chine noise. However, for the case of **human noise** like babble noise and exhibition hall noise, the FB system can obtain a better performance than the FST system.

3. When the recognition results beyond the **10 dB** dataset (**5 dB**, **0 dB**, **-5 dB**) are ignored, the FST system with the MCE-derived SB weights performs slightly better than the FB system. The same results can be found in the statistical significance tests in Appendix D. It is found that the FST system is an effective system to deal with the dataset with a high signal-to-noise ratio. For low signal-to-noise ratio datasets, the FST system performs differently. It can maintain a similar performance for the cases of car noise and subway noise but a degraded performance for the cases of babble nose and exhibition hall noise.

5.5 Experimental III : Full-Band System with Sub-band Energy under Multi-Condition Training

5.5.1 Motivation

The experimental results in Chapter 5.4 have reported that the recognition result of the FST system for the **clean** dataset is better than that of the FB system. The use of the FST system has led to a decrease of about 27% word error rate relative to the FB system for the **clean** dataset. In addition to simply observing a reduction in word error rate, it is also important to try to understand why such a reduction occur. The hypothesis is the use of the sub-band energy on the FST system. It is hypothesized that the SB energy

has better resolution for modeling the energy of the input speech than the traditional FB energy.

To study the performance gain from the SB energy, a new FB model is trained by incorporating the SB energy. The original FB energy is replaced by the two SB energy terms to form the whole acoustic feature vector. By comparing the results of the FB system and the FB system with SB energy, the effectiveness of incorporating SB energy in a model can be easily evaluated.

5.5.2 Setup

The HMM architecture of the new FB system is exactly the same as the FB system defined in 3.3.2 except the new system has 3 more parameters (3 energy terms) than the original one. The full-band energy term is not included in the acoustic vector, the new acoustic features has 42 parameters which is shown below:

- Number of MFCCs : 12 MFCCs + SB1 Energy + SB2 Energy + delta + acceleration

The multi-condition training dataset is used as the training dataset for training the new FB model. The training procedures of the new FB system is the same as the original FB system in order to have a fair comparison between the two systems.

5.5.3 Experimental Result

Figures 5.11 - 5.14 shows the average recognition results between the FST system, FST system with SB energy, FST system with MCE-derived sub-band weights for each noise. The presentation of the recognitions results is

the same as that in Section 5.4.3.

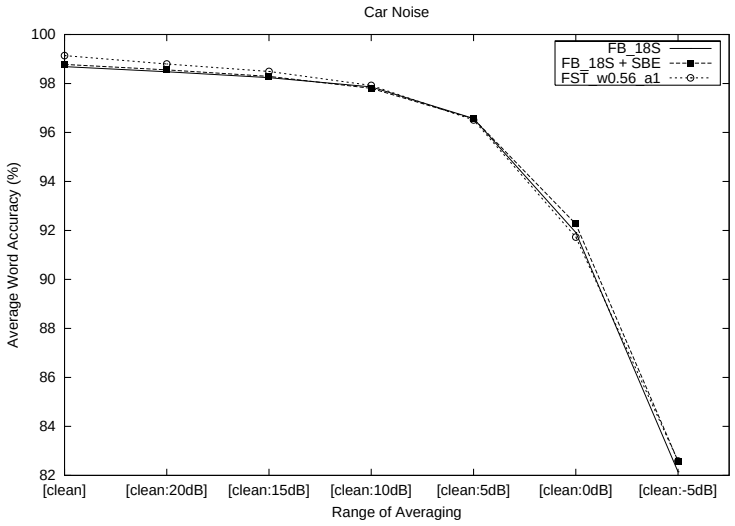

Figure 5.11: Comparison of the recognition results between the FB system, FB system with SB energy and FST system with MCE-derived SB weights on car noise under **multi-condition training** (**FB_18S** = 18-state FB system, **FB_18S + SBE** = 18-state FB system with SB energy, **FST_w0.56_a1** = FST system with MCE-derived sub-band weights [**0.56:0.44**] and asynchrony limit of one state)

Table 5.7 gives a summary for the word error rate for all the systems on the **clean** dataset. **FB 18S + SBE** refers to the 18-state full-band system with the incorporation of the sub-band energy, while **FST + MCE SB weights** refers to the frequency-stream-tying system with the MCE-derived sub-band weights. The statistical significance tests in Appendix D show that the FST system and the 16-state FB system with the sub-band energy are statistically significant better than the other systems.

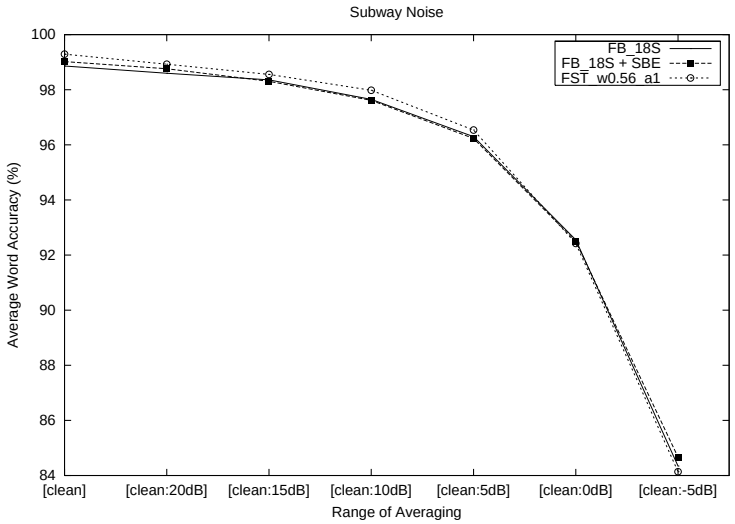

Figure 5.12: Comparison of the recognition results between the FB system, FB system with SB energy and FST system with MCE-derived SB weights on subway noise under **multi-condition training**

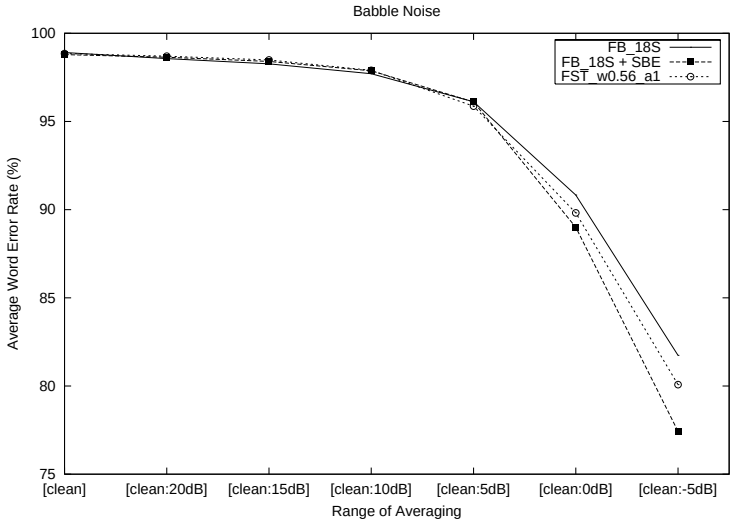

Figure 5.13: Comparison of the recognition results between the FB system, FB system with SB energy and FST system with MCE-derived SB weights on babble noise under **multi-condition training**

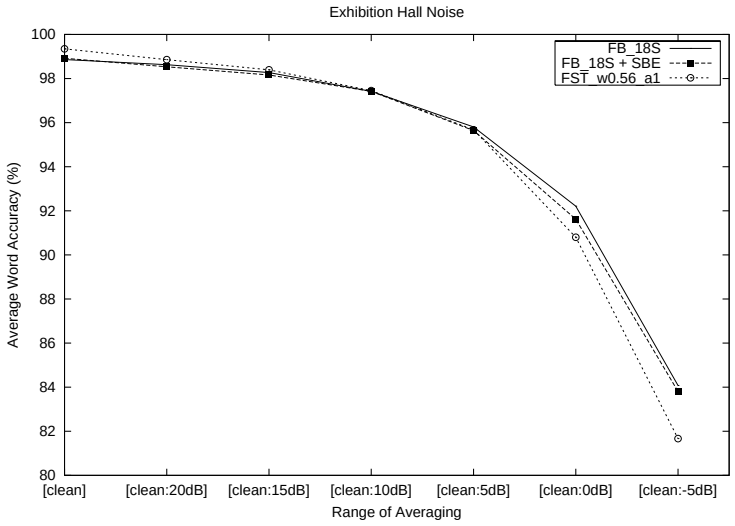

Figure 5.14: Comparison of the recognition results between the FB system, FB system with SB energy and FST system with MCE-derived SB weights on exhibition hall noise under **multi-condition training**

Table 5.7: The word error rate of the frequency-stream-tying system and the full-band system tested on four **clean** datasets under **multi-condition training**, with the % reduction relative to the 18-state FB system

system	**word error rate**	**% reduction**
FB 16S	1.37	+17.1%
FB 18S	1.17	–
FB 18S + SBE	1.12	-4.3%
FB 16S + SBE	0.96	-17.9%
FST	0.87	-25.6%
FST + MCE SB weights	0.84	-28.2%

5.5.4 Discussions

From the results of the FB system with sub-band energy incorporated, it is observed that:

- When concentrating on the recognition result of the **clean** dataset, the performance gain for the 18-state FB system with SB energy over the respective 18-state FB system is small. However, incorporating the SB energy for the 16-state FB system leads to a decrease of about 18% relative error rate. With the addition of the large reduction in the error rate of the FST system, the SB energy is believed to have a better resolution for modeling the energy of the input speech than the traditional FB energy on clean data.

- Among the Figures 5.11 - 5.14, it show that when the recognition result beyond the dataset **10 dB** is ignored, the FST system with the MCE-derived SB weights performs the best.

- The recognition result of the car noise and subway noise show there is a slight improvement or similar performance for the FB system with the incorporation of SB energy. However, for the cases of the babble noise and exhibition hall noise, the performance of the FB system with SB energy degrades significantly for the low signal-to-noise ratio dataset (**0 dB** dataset, **-5 dB** dataset). The poor recognition results help to explain the reason for the poor performances of the FST system on these noises. It is indicated that the sub-band energy is not an appropriate parameter for modeling **human noise** under noisy environments.

5.6 Summary

To be concluded, the FST system can be improved by shifting the band partition to an optimal point, limiting the system to appropriate state asynchrony limit, and optimizing the sub-band weights by MCE training. The performance gains of the FST system from each issue under multi-condition training are summarized in Figure 5.15.

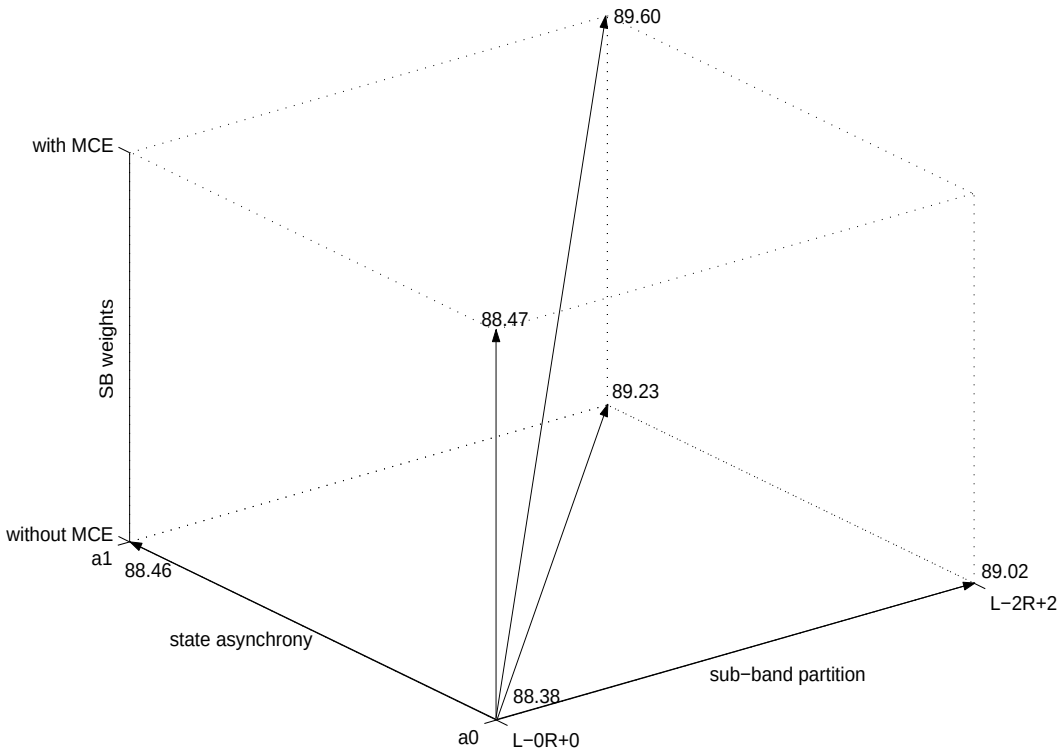

Figure 5.15: Summary of the recognition results for the FST system under **multi-condition training**, with the combinations of shifting to optimal band partition (**sub-band partition**), limiting asynchrony of the sub-band transitions (**state asynchrony**) and optimizing the sub-band weights by MCE training (**SB weights**)

It can be observed that among the three issues, the performance gain from shifting the band partition to an optimal point is the greatest, while the gain from the remaining two issues are similar. The best performance can be achieved by combining all the three issues.

Among the various recognition results of the frequency-stream-tying system and the full-band system, we can conclude:

- The frequency-stream-tying system performs better than the multi-band system. With this new proposed system, the sub-band parameters of the FSTHMM can be jointly optimized through the training process. However, under this frequency-stream-tying HMM framework, the performance gain from the asynchronous recombination of the sub-band acoustic features is smaller than that of the multi-band system. With the use of the optimal band partition on the FSTHMM, it is observed that the system can achieve a satisfactory performance by limiting the asynchrony limit of smaller states, compared to the MB system. The model complexity is thus reduced by the use of the FST system, which can save the computation time of decoding.

- Both the MB system and the FST system outperform the FB system on **clean** dataset. It proposes that the FST systems are more effective under the match-condition (with the FST model trained by the multi-condition training dataset). Under the match-condition, the FST system is statistically significant better than the FB system for the high SNR datasets.

- The FST system with MCE-derived sub-band weights can further improve the performance of the FST system. However, the re-training of the SB weights under the MCE criterion can only shift the weights for a

small value. With that limitation, the FST system with MCE-derived SB weights can only improve the system slightly.

- The FST system with MCE-derived SB weights always perform better than the FB system for the data with a high SNR (excluding **5 dB**, **0 dB** and **-5 dB**). For the case of **machine noise**, like car noise and subway noise, the FST system can achieve similar performance compared to the FB system on the overall. For the case of **human noise** like babble noise and exhibition hall noise, the FB system performs better than the FST system for the data with a low SNR, which results in a better performance on the overall. The statistical significance tests in Appendix D show that there is no statistically significant difference in performance for the FST system with MCE-derived SB weights and the 16-state FB system. However, the 18-state FB system performs the best on the overall.

- The experiment for the FB system with the incorporation of the SB energy proposes that FST system performs better than the FB system for the **clean** dataset is due to the use of the SB energy. The SB energy is indicated to have a better resolution for modeling speech data than FB energy. However, the poor results for the low SNR datasets of the **human noise** propose that the SB energy is not appropriate for modeling the speech consists of **human noise**.

Chapter 6

N-Best Rescoring for Combining the FST System and FB System

6.1 Motivation

It has been shown that under the match-condition between the training and testing environment, the performance of the FST system is similar to that of the FB system on test data with a high SNR, and slightly better in **machine noise** like car and subway noise. In general, two systems may be combined to give better performance provided that the two systems make different errors [30, 31]. Thus, if the FST system and the FB system have different error characteristics, the combined decision from the two systems should lead to a decrease in the system error rate, and should perform at least as well as either of the systems.

In this chapter, the FST system and FB system will be combined by the use of the N-Best rescoring. Normally, the N-Best algorithm [32] is being

used for reducing the search space for the decoding process. For a two-pass search algorithm, the first system is used for generating the N most likely hypotheses, while the second system and usually the better one is used to rescore these N hypotheses. Although the search space has been reduced by the first pass decoding, the two-pass search algorithm can perform as good as the second (better) system given the correct label is included inside the N hypotheses.

However, the N-Best rescoring used in our thesis is not for reducing the search space for the decoding process. It is being used to approximate the combination of the two systems. We use the better system to generate the N hypotheses, while the second system is used to rescore the N hypotheses. The likelihoods of the N hypotheses from the two systems is then summed by the weighting scheme. With this approach, the FB system and the FST system can be combined in a simpler method.

6.1.1 Analysis of the Error Pattern between the FST System and FB System

The assumption made for combining the FST system and FB system is that the two systems have different error characteristics, which leads to different error patterns for the two systems.

In order to check the validity of the assumption, an analysis of the two systems is performed as follows:

1. For each dataset, perform decoding for both the FST system and FB system.

2. For each testing utterance, compares the recognition hypotheses with the true labels.

3. Count the number of utterances that

   (1) FB system recognizes correctly but FST system recognizes incorrectly

   (2) FST system recognizes correctly but FB system recognizes incorrectly

   (3) Both FB system and FST system recognizes incorrectly

Let e_{FB}, e_{FST}, $e_{incorrect}$ be the percentage for FB to make a error, the percentage for the FST system to make a error, the percentage for both two systems to make a error, respectively. The error pattern similarity e_{sim} is defined as:

$$e_{sim} = \frac{e_{incorrect}}{\min\left(e_{FB}, e_{FST}\right)} \tag{6.1}$$

where $\min(a, b)$ is the minimum function which chooses the minimum value for scalar a and b. e_{sim} measures the similarity of the error pattern between the two systems. When e_{sim} equals 1, the error pattern of the two systems will be exactly the same.

Table 6.1 shows the error pattern similarities for the 4 different noises. For each noise, the 7 datasets are divided into 2 subsets: test data with low a SNR and test data with a high SNR. The high SNR datasets include the **clean**, **20 dB**, **15 dB** and **10 dB** datasets, while the remaining datasets are classified into the low SNR datasets. The reason for this division is that all datasets contain a subset have similar error pattern similarities.

It can be observed that the error pattern similarity for the four noises are similar to each other. The error pattern similarity of high SNR datasets are smaller than that of the low SNR datasets. The highly matched error patterns for the low SNR datasets indicate that the error made by the two systems are very similar to each other under noisy environments. It also

Table 6.1: Comparison of the error pattern similarity e_{sim} for the four noises (**Low SNR datasets**: **5 dB**, **0 dB**, **-5 dB**; **High SNR datasets**: **clean**, **20 dB**, **15 dB**, **10 dB**; **Overall** = overall results over all 7 datasets)

Noise Type	High SNR datasets	Low SNR datasets	Overall
Car	0.73	0.87	0.79
Subway	0.66	0.82	0.73
Babble	0.64	0.91	0.76
Exhibition	0.68	0.89	0.77

pointed out that the merged system with the use of N-Best rescoring will not be effective for these datasets. For high SNR datasets, the smaller values of the error pattern similarity point out that the N-Best rescoring scheme may help to improve the performance of the merged system.

6.2 N-Best Rescoring

6.2.1 Methodology

In this chapter, the FST system and FB system will be combined by the use of N-Best Rescoring. The better system is chosen to be the first system to generate the N-Best hypotheses, while the second system will be used to rescore the N-Best sequences. The log-likelihoods of the two systems were summed according to the weighting schemes. The procedures are shown as follows:

1. For each dataset (4 types of noises with 7 types of SNR), by comparing the recognition result of the two systems, the system which performs

better is chosen to be the **1st** system while the remaining one will be the **2nd** system.

2. The N-Best sequences are generated by the **1st** system.

3. For each hypothesis in the N-Best sequences (generated by the **1st** system), the hypothesis is rescored using the **2nd** system.

4. The log-likelihoods from the two systems are summed with the weighting schemes.

5. The N-Best sequences are sorted according to the log-likelihoods generated by the combined systems. The hypothesis with the highest likelihood is chosen to be the recognized string from the combined system.

6.2.2 Related Works

Mirghafori *et al.* [31] have incorporated the FB information streams for their MB system. They have merged the probability streams by simply multiplying the likelihoods from each system and experiments have shown that this can improve the speech recognition accuracy of natural numbers on clean speech data. They have also included an analysis of error patterns of the two systems to explore the reasons for error reduction on the merged system. Besides the above method, other technique, such as boosting, is also used for combining the decisions from different recognizers/classifiers [33].

6.3 Experiment

6.3.1 Setup

Model Selection

As stated before, the FST system performs better under the match-condition between the training and testing environment. In the following experiments, the FSTHMMs are trained under multi-condition training. The FSTHMM with a maximum asynchrony limit of 1 state, and the MCE-derived SB weights [**0.56:0.44**] and the 18-state FB HMM are chosen to generate the N-Best hypotheses for each system.

Number of Hypotheses in N-Best List

In the experiments, the system with the better performance is chosen to be the first system to generate the N-Best hypotheses. The parameter N, which is the number of hypotheses in the N-Best list, does not necessary have to be large. In our experiments, N ranges from 1 to 200. By computing the N hypotheses' combined likelihoods, the best N can be found by comparing the recognition results for different setting of N.

Likelihoods Recombination

The N-Best rescoring technique for combining the FST system and the FB system follows the procedures mentioned in Section 6.2.1 with the log-likelihoods of the two systems are summed with the equal weighting scheme [**0.5:0.5**].

Since the dynamic range of the likelihoods from the two systems are different, direct combination of the log-likelihoods linearly will not be appropriate. The system with a larger dynamic likelihood range will dominate the

combined log-likelihoods and, therefore, make the decisions from the merged system bias to that system. To avoid such phenomenon from happening, the log-likelihoods from each system are normalized by its mean and variance of the N-Best hypotheses, as shown in the following steps:

1. For each utterance in the testing dataset, compute the mean, variance on the N-Best hypotheses' log-likelihoods for the two systems

2. By comparing the mean of the log-likelihoods of the two systems, the system with the smaller mean is chosen for normalization.

3. The N-Best hypotheses' log-likelihoods from the system with the smaller mean are normalized by the mean and variance from the system with the higher mean.

4. After normalization, the log-likelihoods of the two systems will have the same mean and variance.

6.3.2 Experimental Results

In the experiments, N is set to a global value for all datasets. Through the searching of N, it is found to be 90.

Figures 6.1 - 6.4 compare the average word accuracy between the merged system and the individual systems for each noise. The [**dataset1:dataset2**] in the x-axis refers to the recognition result that is averaged from the **dataset1** dataset and **dataset2** dataset, which are the same as the previous chapter.

6.3.3 Discussions

Among the results, several points are worth discussing:

Table 6.2: Comparison of the average word accuracy between the merged system using the N-Best rescoring and the two individual systems under **multi-condition training**

Noise Type	N-Best Rescoring	Individual System	
	N_{global}	**FB_18S**	**FST_w0.56_a1**
Car	90.77	90.56	90.25
Subway	92.00	91.29	91.05
Babble	89.47	89.21	88.01
Exhibition	91.00	90.88	89.09
Average	90.81	90.49	89.60

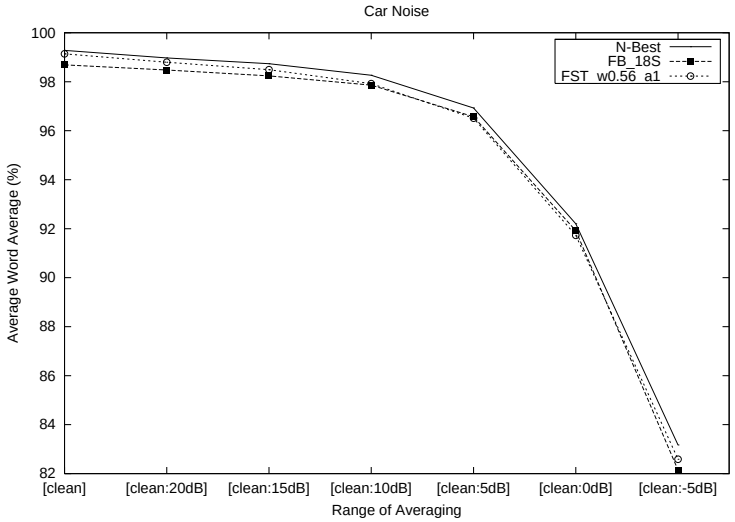

Figure 6.1: Comparison of the recognition results between the merged system using N-Best rescoring and the two individual systems for the car noise under **multi-condition training**

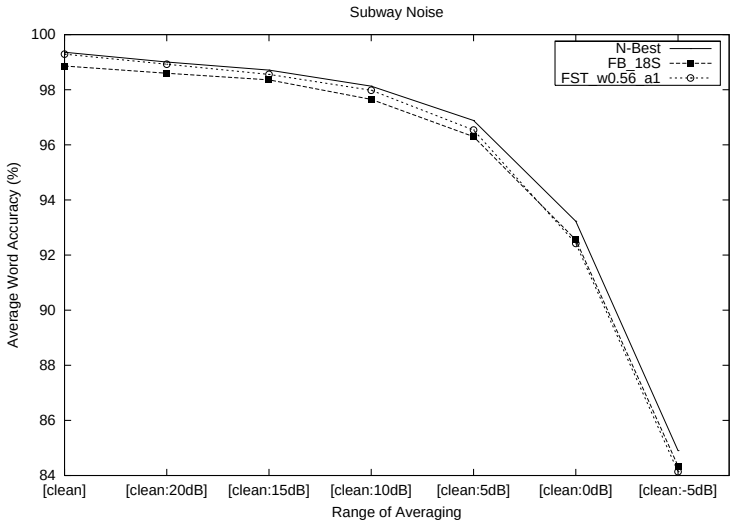

Figure 6.2: Comparison of the recognition results between the merged system using N-Best rescoring and the two individual systems for the subway noise under **multi-condition training**

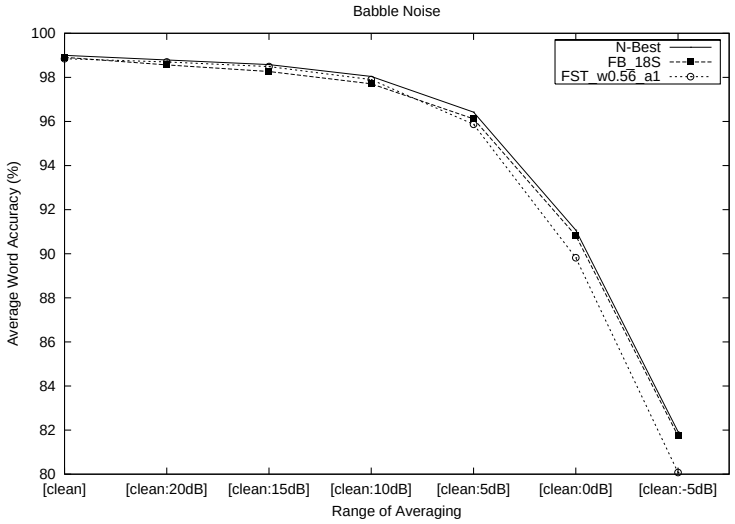

Figure 6.3: Comparison of the recognition results between the merged system using N-Best rescoring and the two individual systems for the babble noise under **multi-condition training**

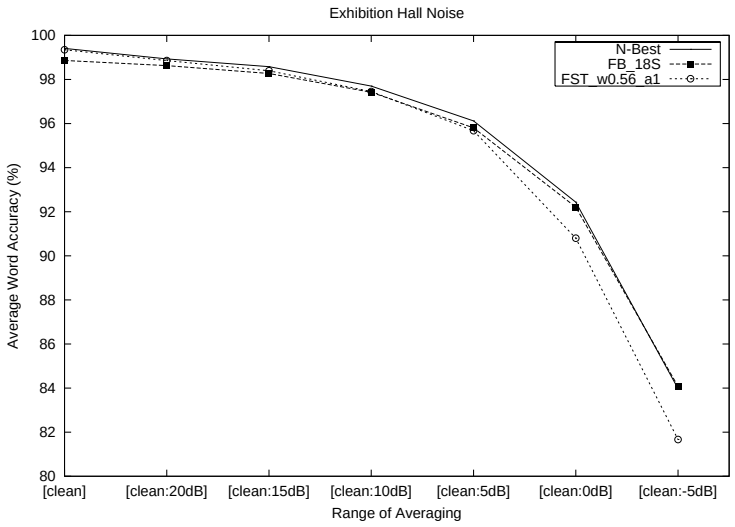

Figure 6.4: Comparison of the recognition results between the merged system using N-Best rescoring and the two individual systems for the exhibition hall noise under **multi-condition training**

- The performance gain of the merged system varies for different noises. Under the subway noise environment, the use of N-Best rescoring on the two systems can further improve about 0.7% over the FB system (which performs slightly better than the FST system). It matches with the error pattern similarity (e_{sim}) measured in the previous section that e_{sim} of the subway noise is the smallest among the four noises. However, for the other noises, the gain is only about 0.3%. The overall performance gain over the four noises is about 0.3% absolutely.

- For all noises, the FST system will always perform better than the FB system for the data with a high SNR and the performance of the merged system is as good as that of the FST system. For the data with a low SNR on car noise or subway noise, the performance of the two systems are similar, which leads to an improvement on the merged system's

performance. However, for the cases of babble noise and exhibition hall noise under low SNR, it can be observed that the FB system performs better than the FST system. Under this situation, the merged system can only gives slight improvement over to the FB system.

6.4 Summary

Based on the experimental results, we can conclude that:

1. The use of the N-Best rescoring scheme to combine the FST system and the FB system can further improve the system's performance. However, the performance gain is small. The main reason is that the error pattern of the two systems are similar to each other.

2. When the error pattern of the two systems are different, the N-Best rescoring scheme can help to improve the performance of the merged system. Among the four noises, the error pattern similarity of the subway noise is the smallest. It matched with the experimental results that the performance gain of the subway noise is the largest among the 4 different noises.

Chapter 7

Conclusion and Future Work

In this chapter, we will summarize the work completed in this thesis and its contributions to the ASR community. In addition, we will also suggest possible future extensions to our work.

7.1 Contributions and Conclusions

7.1.1 Contributions

Our main contribution is the new acoustic modeling method called frequency-stream-tying hidden Markov model (FSTHMM). Under the FSTHMM framework, the three main highlights are:

- **Joint optimization of sub-band parameters through training**
  The traditional MB system separates the acoustic signal into sub-bands and a sub-band model is trained for each sub-band. As a result, the sub-band models can only be optimized individually and independently. In this thesis, the FSTHMM, which we proposed, has multiple streams of acoustic features. Each stream corresponds to the the acoustic features

extracted from one sub-band. As a result, the sub-band parameters of the FSTHMM can be jointly optimized through the training process. The sub-band weights can be trained under the MCE criterion afterward to further improve the system's performance.

- **Asynchronous transitions of sub-bands**

  Researchers hypothesized that the sub-band transitions can occur asynchronously and it may result in a more accurate modeling of the speech signal. With the motivation of the MB model generated by HMM composition, we have designed an FSTHMM with the topology equals to that of the MB HMM. The streams of the FSTHMM are tied across the states as in an MB HMM created by the HMM composition and it can allow asynchronous transitions of different sub-band acoustic features.

- **Stream tying HMM**

  For an FSTHMM, the state space of the model will be much larger than that of the FB HMM when its asynchrony limit is increased. However, through the tying of the frequency streams across the states of the FSTHMM, its number of model parameters can remain nearly the same as that of the FB HMM. The FSTHMM can be viewed as an effective modeling technique for expanding the state space without increasing the number of model parameters.

To further improve the performance of the FST system, we have investigated the following issues in this thesis:

1. **Optimal frequency band partition**

   Under noisy environments, the traditional equal-band partition (64 Hz - 1271 Hz, 1121 Hz - 4000 Hz) is no longer an optimal partition. It should rather shift towards the lower sub-band. Our experimental results show

that the optimal frequency band partition for the 4 noisy environments is somewhere around the band partition **L-2R+2** (64 Hz - 983 Hz, 855 Hz - 4000 Hz).

2. **Allowing state asynchrony**

   We investigated the effect of state asynchrony by training FSTHMMs with different maximum state asynchrony limits. The experimental results show that by allowing a maximum asynchrony of 1 state, our FST system can yield the best performance. Further relaxing asynchrony may result in degraded performance.

3. **Optimization of sub-band weights**

   The sub-band weights can be trained by using the MCE criterion after the training process. However, due to the training procedures of the FST system, the sub-band weights can only be shifted by a small value. The performance gain after incorporating the MCE-derived sub-band weighting scheme is thus small.

4. **Recombination of FST system and FB system**

   The FST system can be merged with the FB system to further improve the system's performance. With the use of the N-Best rescoring scheme, we combined the log-likelihoods of the N-Best hypotheses generated by the two systems. It is shown that the combined system can achieve a slight improvement over the individual systems. This is because the system's error pattern are highly similar.

Table 7.1 - 7.2 summarize the average performance of the FST system with the incorporation of different issues, compared with the FB systems under clean-condition training and multi-condition training, respectively. The

statistical significance tests in Appendix D show that for both the two training conditions, the best FST system is statistically significant better than the best MB system. However, the 18-state FB system performs the best on the overall.

Table 7.1: Comparison of the average word accuracy (over four noises) between the FST system, MB system and the FB system under **clean-condition training** (**MB/FST** = MB/FST system without optimal band partition and state asynchrony; **FB 16/18S** = FB system with 16/18 states; **OBP** = with optimal band partition **L-2R+2**; **Async** = allow state asynchrony)

System	**Average Word Accuracy**
MB	51.93
FST	54.76
FB 16S	56.83
FST + OBP	59.94
MB + OBP + Async	59.98
FST + OBP + Async	60.61
FB 18S	63.38

7.1.2 Conclusions

We have compared the performance of the FSTHMMs with the MB HMMs generated by HMM composition, and also the traditional FB system.

- The experiments on the Aurora2 corpus showed that our FSTHMMs outperform their MB HMM counterparts. Although the performance

Table 7.2: Comparison of the average word accuracy (over four noises) between the FST system, MB system and the FB system under **multi-condition training** (**FST** = FST system without optimal band partition, without state asynchrony, without MCE-derived sub-band weights; **FB 16/18S** = FB system with 16/18 states; **OBP** = with optimal band partition **L-2R+2**; **Async** = allow state asynchrony; **MCE** = with MCE-derived sub-band weights; **Merged System** = merged system with the use of N-Best rescoring)

System	**Average Word Accuracy**
MB + OBP + Async	88.38
FST	88.38
FST + OBP	89.02
FST + OBP + Async	89.23
FST + OBP + Async + MCE	89.60
FB 16S	89.72
FB 18S	90.49
Merged System (FST + FB)	90.83

gain from the asynchronous recombination of sub-band acoustic features is smaller than that of the MB HMM, it is observed that the FST system can achieve the best performance by limiting the asynchrony limit to 1 state under the optimal band partition. Since the model size is smaller than the best MB HMM (with asynchrony limit of two state), the model complexity and computation time is reduced under our FSTHMM framework.

- To make the comparison with the FB system, we compared the 16-state FSTHMMs with two FB HMMs, 16-state and 18-state FB HMM. Under clean-condition training, the FST system has a superior performance over the two FB systems on clean data. This shows that our FSTHMMs should be applied to match condition. Under the multi-condition training method of Aurora2, in which the environments of the training and testing are matched, the FSTHMMs outperform the 16-state FB HMM on car noise and subway noise. The statistical significance tests suggest that there is no statistically significant difference in performance between the FST system and the 16-state FB system on the overall of four noises.

- When the number of states of the FB system is increased to 18, the FSTHMMs give a slight improvement over the 18-state FB HMM on the test data with high SNR. The statistical significance tests show that the improvement is statistically significant. For data with a low SNR, the performance of FSTHMMs varies under different noise environments. It can maintains a similar performance for the car noise and subway noise, which is grouped as **machine noise**, but the performance degraded for the **human noise** like babble noise and exhibition hall noise. Our

experiments on incorporating the sub-band energy on the full-band system showed SB energy can have a better resolution for modeling speech data than the conventional FB energy. However, the poor result of the FSTHMMs and the FB system with the incorporation of SB energy on data with babble noise and exhibition hall noise indicated that the SB energy is not effective for dealing with the **human noise**.

7.2 Future Work

Possible extensions to our work include:

- **Finding the optimal band partition automatically**
  The optimal band partition of the FST system is found by experiments in this thesis. Since the optimal band partition varies for different noises, the performance of the FST systems can be further improved if the optimal band partition can be found automatically. The possible solution maybe achieved by trying to determine the relationship between entropy, sub-band SNR and system performance with different band partition configurations.

- **Improve the performance of FST system on low SNR data**
  The performance of the FST system cannot be as good as the FB system for the data with low SNR, especially for the **human noise** like babble noise and exhibition hall noise. It will be an interesting issue if the FST system can be improved on low SNR data.

- **Improve the merging technique of the two systems**
  Merged the systems by the use of the N-Best rescoring can only improve the system's performance slightly. Besides the highly similar

error patterns between the two systems, the slight improvement may also be due to the direct summation of the likelihoods of the two systems is not effective. If more information such as the alignments of the digit sequences for each hypothesis can be incorporated in the merging scheme, then the merged system would improve significantly.

- **Improve the stream weights**

  The stream weights (SB weights) is initially set as equal to pass the training process. The assumption of equal contribution between sub-bands may not be valid. If the weights can be estimated first, it is expected that the FST system can be further improved.

- **Incorporation of model-dependent weights**

  The sub-bands for each model may have different reliability even under the same noisy environments. It will be helpful to estimate the weights for individual sub-bands instead of using a global sub-band weight.

- **Evaluation on large-vocabulary corpora**

  It will be interesting to evaluate the FST framework on large-vocabulary corpora like phone recognition.

Appendix A

Aurora Baseline Results

The corpus used in this thesis is the Aurora2 corpus, which includes its baseline results. It is thus interesting to compare the 16-state, 18-state FB baseline results used in this thesis with the Aurora2 baseline.

The configurations of the feature extraction in this thesis are not exactly the same compared to the Aurora2 baseline system [20]. The configurations for generating the MFCCs in this thesis include energy normalization for computing the energy terms and zero-mean the speech data, which are not included in Aurora2 baseline configurations. Tables A.1 - A.2 compare the performance of the FB systems used in this thesis and the Aurora2 FB baseline.

It is observed that besides the 16-state FB system trained from clean-condition training data, which gives poor recognition results, all the FB systems in this thesis outperform the FB system used in the Aurora2 baseline. Thus it shows that our FB baseline systems are comparable with the FB system from other researchers.

Table A.1: Average word accuracy of the Aurora2 baseline system and the full-band baseline system under **clean-condition training** (**Aurora2** = baseline result from the Aurora2 corpus)

Noise Type	**Aurora2**	**FB_16S**	**FB_18S**
Car	60.60	56.76	59.03
Subway	69.48	58.69	66.93
Babble	49.88	51.35	59.43
Exhibition	65.39	60.53	68.12
Average	61.34	56.83	63.38

Table A.2: Average word accuracy of the Aurora2 baseline system and the full-band baseline system **multi-condition training**

Noise Type	**Aurora2**	**FB_16S**	**FB_18S**
Car	86.52	89.73	90.56
Subway	88.75	90.14	91.29
Babble	87.95	88.83	89.21
Exhibition	88.03	90.18	90.88
Average	87.81	89.72	90.49

Appendix B

16-state Multi-Band HMM vs 18-state Multi-Band HMM

The multi-band HMM generated by HMM composition always has more states than the full-band system. In order to have a fair comparison between the FB system and the MB/FST system, the number of states of the FB HMM is increased so as to determine the best FB system. In the full-band baseline system, it can be observed that increasing the number of states of the FB HMM from 16 states to 18 states can yield a performance gain, while further increasing the states will harm the performance. The performance of the 18-state FB system is thus chosen as the FB baseline system to compare with that of the 16-state MB/FST system. Such comparison will give rise to the question: Will the MB/FST system improve when the number of states of its HMM is also increased?

For simplicity, only the MB system under clean-condition training is chosen for the investigation of the effect of increasing number of states on its HMMs. The configurations and the environment settings of the 18-state MB HMMs are exactly the same as the 16-state MB HMMs. Figures B.1 shows

the performance of the 18-state MB system with different band partition settings.

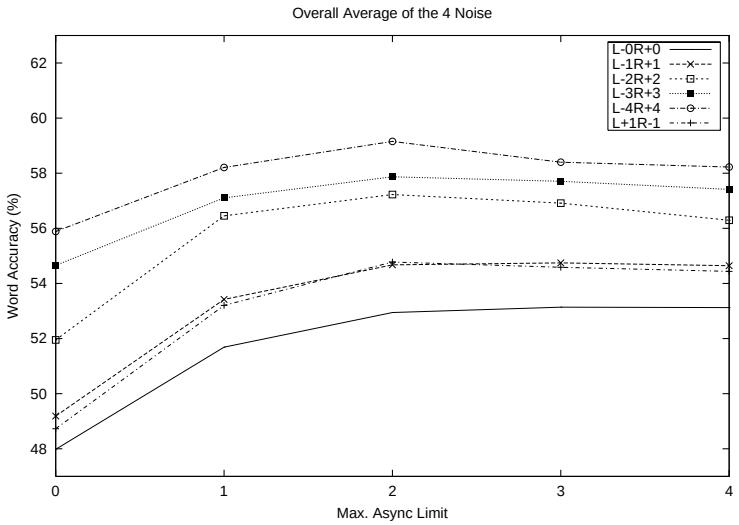

Figure B.1: Average recognition results for the 18-state MB system on four noises under **clean-condition training**, with maximum asynchrony limit up to four states

It can be observed from the figure that the 18-state multi-band system performs the best with the use of band partition **L-4R+4**. The best result can be achieved by limiting the asynchrony limit to two states. Table B.1 compares the performance of the best 18-state multi-band system (using the band partition **L-4R+4**) with the best 16-state MB system (using the band partition **L-2R+2**).

Unlike the FB system, the 18-state MB system does not give a performance gain over the 16-state system. The performance of the two systems vary under different noise environments, but the 16-state MB system performs slightly better overall. Although the performances of the 18-state FST system or the system under multi-condition training are not known, it is

Table B.1: Recognition results for the best 18-state multi-band system and the best 16-state multi-band system on four noises under **clean-condition training**

	Multi-Band System	
	18-state MB system	**16-state MB system**
Noise	**L-4R+4, a2**	**L-2R+2, a2**
Car	56.17	58.38
Subway	65.16	63.36
Babble	54.55	58.30
Exhibition	60.74	59.86
Average	59.16	59.98

expected that the two systems will have a similar performance. In conclusion, the 16-state MB/FST system can be viewed as the best system to be compared with the FB system.

Appendix C

Detailed Experimental Results

C.1 Clean-Condition Training

Table C.1: Recognition results for the best MB system **MB_w0.5_a2** on four noises under **clean-condition training** in Table 7.1

dataset	Car	Subway	Babble	Exhibition	Average
clean	99.31	99.36	98.91	99.54	99.28
20 dB	97.26	95.12	97.31	95.87	96.39
15 dB	89.26	88.55	85.22	88.46	87.87
10 dB	62.72	71.91	59.46	68.03	65.53
5 dB	30.87	41.48	34.98	33.79	35.28
0 dB	11.81	19.74	14.51	13.17	14.81
-5 dB	7.40	9.49	7.80	8.52	8.30
Average	58.38	63.36	58.30	59.86	59.98

Table C.2: Recognition results for the best FST **FST_w0.5_a1** system on four noises under **clean-condition training** in Table 7.1

dataset	**Car**	**Subway**	**Babble**	**Exhibition**	**Average**
clean	99.49	99.48	99.06	99.63	99.42
20 dB	97.02	96.41	94.86	95.56	95.96
15 dB	86.07	90.11	83.04	87.94	86.79
10 dB	66.12	75.31	60.58	69.58	67.90
5 dB	33.10	46.12	35.28	38.82	38.33
0 dB	11.60	16.61	14.18	13.95	14.09
-5 dB	7.58	8.90	8.52	8.18	8.29
Average	58.78	64.91	57.59	61.17	60.61

Table C.3: Recognition results for the 18-state FB system **FS_18S** on four noises under **clean-condition training** in Table 7.1

dataset	**Car**	**Subway**	**Babble**	**Exhibition**	**Average**
clean	99.16	99.20	99.09	99.38	99.21
20 dB	98.42	97.85	97.82	97.38	97.87
15 dB	93.56	94.47	88.51	94.66	92.80
10 dB	59.41	77.92	57.41	83.71	69.61
5 dB	29.67	43.38	35.61	45.11	38.44
0 dB	14.08	21.03	17.78	19.72	18.15
-5 dB	9.22	10.53	8.95	11.79	10.12
Average	59.03	66.93	59.43	68.12	63.38

C.2 Multi-Condition Training

Table C.4: Recognition results for the best MB system **MB_w0.5_a4** on four noises under **multi-condition training** in Table 7.2

dataset	Car	Subway	Babble	Exhibition	Average
clean	98.93	98.86	98.91	99.23	98.98
20 dB	98.09	97.97	98.31	97.75	98.03
15 dB	97.49	97.51	97.67	96.58	97.31
10 dB	95.71	94.96	95.95	93.58	95.05
5 dB	89.38	87.69	87.52	85.65	87.56
0 dB	64.81	66.84	60.43	63.71	63.95
-5 dB	23.62	30.64	18.68	24.53	24.37
Average	89.10	88.99	87.98	87.45	88.38

Table C.5: Recognition results for the best FST system **FST_w0.56_a1** on four noises under **multi-condition training** in Table 7.2

dataset	Car	Subway	Babble	Exhibition	Average
clean	99.14	99.29	98.85	99.35	99.16
20 dB	98.45	98.56	98.55	98.36	98.48
15 dB	97.88	97.82	98.07	97.47	97.81
10 dB	96.21	96.25	96.13	94.63	95.81
5 dB	90.87	90.76	87.76	88.58	89.49
0 dB	67.82	71.85	59.55	66.40	66.41
-5 dB	27.74	34.42	21.64	26.87	27.67
Average	90.25	91.05	88.01	89.09	89.60

Table C.6: Recognition results for the 16-state FB system **FB_16S** on four noises under **multi-condition training** in Table 7.2

dataset	Car	Subway	Babble	Exhibition	Average
clean	98.45	98.68	98.76	98.64	98.63
20 dB	97.94	97.94	98.00	97.72	97.90
15 dB	97.52	97.11	97.64	97.22	97.37
10 dB	96.33	95.39	96.04	94.75	95.63
5 dB	90.34	89.25	88.88	88.43	89.22
0 dB	66.54	71.02	63.57	72.76	68.47
-5 dB	24.13	30.43	29.05	33.60	29.30
Average	89.73	90.14	88.83	90.18	89.72

Table C.7: Recognition results for the 18-state FB system **FB_18S** on four noises under **multi-condition training** in Table 7.2

dataset	**Car**	**Subway**	**Babble**	**Exhibition**	**Average**
clean	98.69	98.86	98.91	98.86	98.83
20 dB	98.27	98.34	98.22	98.40	98.31
15 dB	97.76	97.88	97.67	97.56	97.72
10 dB	96.72	95.52	96.04	94.85	95.78
5 dB	91.41	90.88	89.75	89.36	90.35
0 dB	68.65	73.84	64.36	74.24	70.27
-5 dB	23.53	35.09	27.21	35.27	30.27
Average	90.56	91.29	89.21	90.88	90.49

Appendix D

Statistical Significance Tests

The statistical significance test from NIST (National Institute of Standard and Technology) has been used within the DARPA speech research community. It can be used to report the empirical results by using statistical measures in summarizing the findings and drawing conclusions. The package includes four tests:

- Match Pairs Sentence Segment (Word Error) Test (MP)
- Signed Paired Comparison (Speaker Word Accuracy Rate (%)) Test (SI)
- Wilcoxon Signed Rank (Speaker Word Accuracy Rate (%)) Test (WI)
- McNemar (Sentence Error) Test (MN)

We applied the tests to compare the performance of the various systems under four conditions:

1. **Overall performance of various systems under clean-condition training**

It compares the overall performance of the 18-state FB system with the best MB system **MB_w0.5_a2** (MB + OBP + Async) and the best FST system **FST_w0.5_a1** (FST + OBP + Async) of Table 7.1. The test results are shown in Table D.1. A system identifier showing up in a table entry means it is the better system when compared with the system having the row identifier. Among the four tests, we focus on the MP Test which compares the overall word error between the systems. It shows that the FST system is significantly better than the best MB system. However, among the three systems, the 18-state FB system performs the best.

2. **Overall performance of various systems under multi-condition training**

   It compares the overall performance between the 16-state FB system, 18-state FB system, best MB system **MB_w0.5_a4** (MB + OBP + Async) and the best FST system **FST_w0.56_a1** (FST + OBP + Async + MCE) of Table 7.2. The test results in Table D.2 point out that the 18-state FB system performs the best among the four systems and there is no statistically significant difference in performance between the 16-state FB system and the best FST system. The test results also show the MB system is not good enough compared with the other systems.

3. **Performance of various systems on clean data under multi-condition training**

   It compares the performance between the 16-state FB system, 16-state FB system with sub-band energy **FB_16S_SBE**, 18-state FB system, 18-state FB system with sub-band energy **FB_18S_SBE** and the FST

system of Table 5.7. The results of the statistical significance tests are shown in Table D.3. It shows that among the five systems, the FST system and the 16-state FB system with the sub-band energy is statically significant better than the other systems on clean data.

4. **Performance of various systems on high SNR data under multi-condition training**

   It compares the performance between the 16-state FB system, 18-state FB system, the best MB system **MB_w0.5_a4** and the best FST system **FST_w0.56_a1** on data with high SNR (Tables C.4 - C.7). The high SNR datasets include the **clean** dataset, **20 dB** dataset, **15 dB** dataset and **10 dB** dataset. The results of the statistical significance tests are shown in Table D.4. It shows that the FST system **FST_w0.56_a1** performs the best and is statistically significant better than the 18-state FB system on data with high SNR.

Table D.1: Statistical significance tests on various systems under clean-condition training based on their overall performance on Aurora2 test A (as shown in Table 7.1)

Composite Report of All Significance Tests
For the Test on the Overall Performance under Clean-Condition Training

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy Rate (%)) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy Rate (%)) Test	WI
McNemar (Sentence Error) Test	MN

	MB_w0.5_a2	FST_w0.5_a1	FB_18S
MB_w0.5_a2		MP FST_w0.5_a1 SI FST_w0.5_a1 WI FST_w0.5_a1 MN same	MP FB_18S SI FB_18S WI FB_18S MN FB_18S
FST_w0.5_a1			MP FB_18S SI FB_18S WI FB_18S MN FB_18S
FB_18S			

Table D.2: Statistical significance tests on various systems under multi-condition training based on their overall performance on Aurora2 test A (as shown in Table 7.2)

Composite Report of All Significance Tests
For the Test on the Overall Performance under Multi-Condition Training

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy Rate (%)) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy Rate (%)) Test	WI
McNemar (Sentence Error) Test	MN

	MB_w0.5_a4	FB_16S	FST_w0.56_a1	FB_18S
MB_w0.5_a4		MP FB_16S SI FB_16S WI FB_16S MN FB_16S	MP FST_w0.56_a1 SI FST_w0.56_a1 WI FST_w0.56_a1 MN FST_w0.56_a1	MP FB_18S SI FB_18S WI FB_18S MN FB_18S
FB_16S			MP same SI FB_16S WI same MN same	MP FB_18S SI FB_18S WI FB_18S MN FB_18S
FST_w0.56_a1				MP FB_18S SI FB_18S WI FB_18S MN FB_18S
FB_18S				

Table D.3: Statistical significance tests on various systems under multi-condition training based on their performance on clean testing data only (as shown in Table 5.7)

Composite Report of All Significance Tests
For the Test on Clean Testing Data under Multi-Condition Training

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy Rate (%)) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy Rate (%)) Test	WI
McNemar (Sentence Error) Test	MN

	FB_16S	FB_18S	FB_18S_SBE	FB_16S_SBE	FST
FB_16S		MP FB_18S SI same WI same MN same	MP FB_18S_SBE SI same WI same MN FB_18S_SBE	MP FB_16S_SBE SI FB_16S_SBE WI FB_16S_SBE MN FB_16S_SBE	MP FST SI FST WI FST MN FST
FB_18S			MP same SI same WI same MN same	MP FB_16S_SBE SI same WI FB_16S_SBE MN FB_16S_SBE	MP FST SI FST WI FST MN FST
FB_18S_SBE				MP FB_16S_SBE SI same WI same MN same	MP FST SI FST WI FST MN FST
FB_16S_SBE					MP same SI same WI same MN same
FST					

Table D.4: Statistical significance tests on various systems under multi-condition training based on their performance on testing data with high SNR: **clean**, **20 dB**, **15 dB** and **10 dB** (as shown in Table C.4 - C.7)

Composite Report of All Significance Tests
For the Test on High SNR Testing Data under Multi-Condition Training

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy Rate (%)) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy Rate (%)) Test	WI
McNemar (Sentence Error) Test	MN

	MB_w0.5_a4	FB_16S	FB_18S	FST_w0.56_a1
MB_w0.5_a4		MP FB_16S SI FB_16S WI same MN same	MP FB_18S SI FB_18S WI FB_18S MN FB_18S	MP FST_w0.56_a1 SI FST_w0.56_a1 WI FST_w0.56_a1 MN FST_w0.56_a1
FB_16S			MP FB_18S SI same WI same MN FB_18S	MP FST_w0.56_a1 SI same WI same MN FST_w0.56_a1
FB_18S				MP FST_w0.56_a1 SI same WI same MN FST_w0.56_a1
FST_w0.56_a1				

Bibliography

[1] J. B. Allen, How Do Humans Process and Recognize Speech. *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 567-577, 1994.

[2] Y. C. Tam, Development of an Asynchronous Multi-Band System for Continuous Speech Recognition. MPhil thesis, The Hong Kong University of Science and Technology, Hong Kong, 1997.

[3] M. J. Tomlinson, M. J. Russell, R. K. Moore, A. P. Buckland and M. A. Fawley. Modeling Asynchrony in Speech Using Elementary Single-signal Decomposition. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, vol. 2, pp. 1247-1250, 1997.

[4] S. Matsuda, M. Nakai, H. Shimodaira and S. Sagayama. Asynchronous-Transition HMM. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 2, pp. 1001-1004, 2000.

[5] H. Bourlard and S. Dupont, A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[6] H. Hermansky, S. Tibrewala, and M. Pavel, Towards ASR on Partially Corrupted Speech. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[7] C. Cerisara, J.P. Haton, J.F. Mari, and D. Fohr, Multi-band Continuous Speech Recognition, In *Proceeding of the European Conference on Speech Communication and Technology*, 1997

[8] S. Okawa, E. Bocchieri, and A. Potamianos, Multi-Band Speech Recognition in Noisy Environment. In *Proceeding of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 2, pp. 641-644, 1998.

[9] S. Okawa, T. Nakajima, and K. Shirai, A Recombination Strategy for Multi-band Speech Recognition Based on Mutual Information Criterion. In *Proceedings of the European Conference on Speech Communication and Technology*, vol. 2, pp. 603-606, 1999.

[10] S. Tibrewala and H. Hermansky, Sub-Band Based Recognition of Noisy Speech, In *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[11] C. Cerisara and D. Fohr, Multi-band Automatic Speech Recognition *Journal of Computer Speech and Language*, vol. 15, issue 2, pp. 151-174, 2001.

[12] A. P. Varga and R. K. Moore, Hidden Markov Model Decomposition of Speech and Noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 845-848, 1990.

[13] N. Mirghafori, A Multi-Band Approach to Automatic Speech Recognition, Ph.D. thesis, University of California at Berkeley, Berkeley, CA., Dec. 1998.

[14] C. Cerisara, D. Fohr and J. P. Haton, Asynchrony in Multi-Band Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1121-1124, 2000.

[15] C. Cerisara, J. P. Haton, and D. Fohr, A Recombination Model for Multi-band Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[16] Y. C. Tam and B. Mak, Optimization of Sub-Band Weights using Stimulated Noisy Speech in Multi-Band Speech Recognition. In *Proceedings of International Conference on Spoken Language Processing*, vol. 1, pp. 313-316, 2000.

[17] Y. C. Tam and B. Mak, An Alternative Approach of Finding Competing Hypotheses for Better Minimum Classification Error Training. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, vol. 1, pp. 101-104, 2002.

[18] P. McCourt, S. Vaseghi and N. Harte. Multi-resolution cepstral features for phoneme recognition across speech sub-bands. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, vol. 1, pp. 557-560, 1998.

[19] A. Hagen and A. Morris, Comparison of HMM Experts with MLP Experts in the Full Combination Multi-Band Approach to Robust ASR.

In *Proceedings of the International Conference on Spoken Language Processing*, 2002.

[20] H.G. Hirsch, D. Pearce, The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition System Under Noisy Conditions. In *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, 2000.

[21] Aurora Evaluations, http://www.isip.msstate.edu/projects/aurora/.

[22] R. G. Leonard, A Database for Speaker-Independent Digit Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1984.

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

[24] B. Mak, Y. C. Tam, Asynchrony with Trained Transition Probabilities Improves Performance in Multi-Band Speech Recognition. In *Proceedings of International Conference on Spoken Language Processing*, vol. 4, pp. 149-152, 2000.

[25] S. Young, D. Kershaw, J. Odell, D. Ollason, V.Valtchev and P. Woodland, The HTK book – version2.2, Entropic, 1999.

[26] B. Logan and P. Moreno, Factorial HMMS for Acoustic Modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 813-816, 1998.

[27] S. Russell and P. Norving, Artificial Intelligence A Modern Approach, Prentice Hall, 1995.

[28] H. Goltin, D. Vergyri, C. Neti, G. Potamianos and J. Luettin, Weighting schemes for audio-visual fusion in speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 173-176, 2001.

[29] M. L. Shire, Multi-Stream ASR Trained with Heterogeneous Reverberant Environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[30] S. L. Wu, B. E. D. Kingsbury, N. Morgan and S. Greenberg, Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721-724, 1998.

[31] N. Mirghafori and N. Morgan, Combining Connectionist Multi-Band and Full-Band Probability Streams for Speech Recognition of Natural Numbers. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 743-746, 1998.

[32] R. Schwartz and Y. L. Chow, The N-Best Algorithm: an Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 81-84, 1990.

[33] H. Schwenk, Using Boosting to Improve a Hybrid HMM/Neural Network Speech Recognizer. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, March 1999.